SPECIAL ISSUE

Analysis and Mining of Social Media Data

A *PeerJ Computer Science* Special Issue highlighting novel research that mines and analyzes data from social media platforms.



PeerJ Computer Science

A *PeerJ Computer Science* Special Issue

Analysis and Mining of Social Media Data

Editor Feature

PeerJ Computer Science

Special Issue Editor



Arkaitz Zubiaga
Senior Lecturer
Associate Professor
Queen Mary University of London
UK

Dr. Arkaitz Zubiaga is a senior lecturer (associate professor) at Queen Mary University of London. His research revolves around Social Data Science, interdisciplinary research bridging NLP and Computational Social Science. He is particularly interested in linking online data with events in the real world, among others for tackling problematic issues on the Web and social media that can have a damaging effect on individuals or society at large, such as hate speech, misinformation and inequality.

"*Social media platforms give us access to datasets of unprecedented scale that enable analysis of a wide range of social phenomena by using state-of-the-art methods for computational social science. There is however also malicious activity in social media that needs countering, as is the case with hate speech and misinformation.***"**

Special Issue Editor



Paolo Rosso
Professor
Universidad Politécnica de Valencia
Spain

Prof. Paolo Rosso is Full Professor at the Universitat Politècnica de València, where he is also a member of the Pattern Recognition and Human Language Technology (PRHLT) research center. His research interests are focused on natural language processing, mainly on fake news and hate speech detection, author profiling, and irony detection. He has published 50+ articles in journals and 400+ articles in conferences and workshops.He is in the ranking of the top H-index scientists in Spain

"Developing tools and methods for mining social media data has become increasingly important in the era of big data, where a large number of users are using social media platforms even through mobile devices."

PeerJ Computer Science

Analysis and Mining of Social Media Data

Special Issue on Analysis and Mining of Social Media Data. Arkaitz Zubiaga, Paolo Rosso. https://doi.org/10.7717/peerj-cs.1909

Using social media advertisement data to monitor the gender gap in STEM: opportunities and challenges. Reham Al Tamime, Ingmar Weber. https://doi.org/10.7717/peerj-cs.994

How Twitter data sampling biases U.S. voter behavior characterizations. Kai-Cheng Yang, Pik-Mai Hui, Filippo Menczer. https://doi.org/10.7717/peerj-cs.1025

RuSentiTweet: a sentiment analysis dataset of general domain tweets in Russian. Sergey Smetanin. https://doi.org/10.7717/peerj-cs.1039

Is gender-based violence a confluence of culture? Empirical evidence from social media. Rimjhim, Sourav Dandapat. https://doi.org/10.7717/peerj-cs.1051

Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. Hind Almerekhi, Haewoon Kwak, Bernard J. Jansen. https://doi.org/10.7717/peerj-cs.1059

An optimized deep learning approach for suicide detection through Arabic tweets. Nadiah A. Baghdadi, Amer Malki, Hossam Magdy Balaha, Yousry AbdulAzeem, Mahmoud Badawy, Mostafa Elhosseini. https://doi.org/10.7717/peerj-cs.1070

COVID-19 mass media coverage in English and public reactions: a West-East comparison via Facebook posts. Ahmad R. Pratama, Firman M. Firmansyah. https://doi.org/10.7717/peerj-cs.1111

Resilience of political leaders and healthcare organizations during COVID-19. Manmeet Kaur Baxi, Joshua Philip, Vijay Mago. https://doi.org/10.7717/peerj-cs.1121

Effect of visual imagery in COVID-19 social media posts on users' perception. Waleed M. Al-nuwaiser. https://doi.org/10.7717/peerj-cs.1153

An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests. Hieu Nguyen, Swapna Gokhale. https://doi.org/10.7717/peerj-cs.1127

Sentiment analysis of vegan related tweets using mutual information for feature selection. Elvina Shamoi, Akniyet Turdybay, Pakizar Shamoi, Iskander Akhmetov, Assel Jaxylykova, Alexandr Pak. https://doi.org/10.7717/peerj-cs.1149

PeerJ Computer Science

Analysis and Mining of Social Media Data

Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations. Rasha Obeidat, Maram Gharaibeh, Malak Abdullah, Yara Alharahsheh. https://doi.org/10.7717/peerj-cs.1151

Addressing religious hate online: from taxonomy creation to automated detection. Alan Ramponi, Benedetta Testa, Sara Tonelli, Elisabetta Jezek. https://doi.org/10.7717/peerj-cs.1128

Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. Dan Heaton, Jeremie Clos, Elena Nichele, Joel Fischer. https://doi.org/10.7717/peerj-cs.1211

Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: a study on governmental preventive measures. Muhammad Faisal Ali, Rabia Irfan, Tahira Anwar Lashari. https://doi.org/10.7717/peerj-cs.1220.

Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets. Ahmad Fathan Hidayatullah, Rosyzie Anna Apong, Daphne T.C. Lai, Atika Qazi. https://doi.org/10.7717/peerj-cs.1312

BERT-based tourism named entity recognition: making use of social media for travel recommendations. Dhomas Hatta Fudholi, Annisa Zahra, Septia Rani, Sheila Nurul Huda, Irving Vitra Paputungan, Zainudin Zukhri. https://doi.org/10.7717/peerj-cs.1731

PeerJ Computer Science

Special issue on analysis and mining of social media data

Arkaitz Zubiaga[1] and Paolo Rosso[2]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom
[2] Technical University of Valencia, Valencia, Spain

This Editorial article has not been externally peer reviewed

ABSTRACT

This Editorial introduces the PeerJ Computer Science Special Issue on Analysis and Mining of Social Media Data. The special issue called for submissions with a primary focus on the use of social media data, for a variety of fields including natural language processing, computational social science, data mining, information retrieval and recommender systems. Of the 48 abstract submissions that were deemed within the scope of the special issue and were invited to submit a full article, 17 were ultimately accepted. These included a diverse set of articles covering, *inter alia*, sentiment analysis, detection and mitigation of online harms, analytical studies focused on societal issues and analysis of images surrounding news. The articles primarily use Twitter, Facebook and Reddit as data sources; English, Arabic, Italian, Russian, Indonesian and Javanese as languages; and over a third of the articles revolve around COVID-19 as the main topic of study. This article discusses the motivation for launching such a special issue and provides an overview of the articles published in the issue.

Subjects Computational Linguistics, Data Science, Network Science and Online Social Networks, Social Computing, World Wide Web and Web Science
Keywords Social media, Data mining, Natural language processing, Computational social science

Submitted 26 January 2024
Accepted 1 February 2024
Published 29 February 2024

Corresponding author
Arkaitz Zubiaga, arkaitz@zubiaga.org

Article type
Editorial

Additional Information and Declarations can be found on page 6

DOI **10.7717/peerj-cs.1909**

Copyright
2024 Zubiaga and Rosso

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

The number of users in online social media has increased substantially in the last two decades. With the growing number of users posting and resharing content, interacting with others and exchanging files, online social media platforms have become an invaluable data source to perform scientific analyses at scale. These analyses are conducted in a wide range of scientific fields, including research in computational social science, understanding social phenomena, natural language processing at scale and social network analysis investigating how interactions are shaped online, *inter alia*.

Along with the many advantages that social media platforms have offered with unprecedented means for communication with one another, they are also often misused for malicious purposes. Indeed, there is evidence of many users exploiting social media platforms to perform uncivil actions and engage in toxic conversations. This can cause harmful consequences for online users, many of whom have no other option but to delete their social media accounts. Detection and mitigation of this kind of activity has become

a priority in the scientific community, as is the need to further analyse, measure and understand this phenomenon.

With these and other related challenges in mind, this special issue sought a diverse set of contributions with novel research analysing and/or mining social media data to address a wide range of priorities in the scientific community, including but not limited to: (i) analysis; detection; resolution and mitigation of online malicious communications, (ii) computational social science; social network analysis and social computing, (iii) measurement; analysis and modeling of social media–including real-time analytics, (iv) data collection and annotation from social media, (v) trust, privacy and security in social media, (vi) natural language processing for social media, (vii) recommender systems for social media, and (viii) information retrieval and search in social media. Topics beyond those listed above were also welcome as long as the fit to the scope of the special issue was clear.

SPECIAL ISSUE THEMES

The special issue received a total of 74 abstract submissions, of which 48 were deemed within the scope of the issue and were invited to submit full articles for review. After peer review, 17 articles were ultimately accepted to be included in the special issue.

Among the published articles, we identify a diverse range of themes which make up the special issue. As many as six articles focus on **sentiment analysis** for different purposes (*Smetanin, 2022*; *Pratama & Firmansyah, 2022*; *Baxi, Philip & Mago, 2022*; *Nguyen & Gokhale, 2022*; *Shamoi et al., 2022*; *Ali, Irfan & Lashari, 2023*). Four studies focused on tackling **online harms** of different kinds, with studies on abusive language detection (*Almerekhi, Kwak & Jansen, 2022*; *Ramponi et al., 2022*), suicidal ideation detection (*Baghdadi et al., 2022*) and misinformation detection (*Obeidat et al., 2022*). Others studied **NLP techniques for social media**, focused on the analysis of Twitter discourse (*Heaton et al., 2023*), language identification (*Hidayatullah et al., 2023*) and named entity recognition (*Fudholi et al., 2023*). There are also a number of **analytical studies** that investigate different societal issues, including gender equality through advertising data (*Al Tamime & Weber, 2022*), analysis of gender-based violence across countries (*Rimjhim & Dandapat, 2022*) as well as electoral data (*Yang, Hui & Menczer, 2022*). Another study looked at the **effect of visual images** on social media news (*Alnuwaiser, 2022*).

All in all, these articles make a diverse contribution to the literature of analysing and mining social media data, highlighting the wide range of research questions that social media as a data source enables us to answer.

Further, looking at the diversity in the published articles, we look at three different dimensions:

- **COVID-19:** given the temporal overlap of the special issue with the pandemic, it attracted six submissions related to COVID-19.

- **Social media platforms:** by far the most common social media platform used as a data source was Twitter, with 13 articles. Two articles use Facebook data, one uses Reddit data and one uses web data.
- **Languages:** while most studies focus on English data collected from social media, a few of the articles use data in other languages. Indeed, there are two articles using Arabic data, one using Italian data, one using Russian data and one using Indonesian and Javanese data.

SUMMARY OF CONTRIBUTIONS

The 17 articles accepted to the special issue include the following, in order of publication:

1. *Al Tamime & Weber (2022)* used advertisement data collected from Facebook to analyse the interests of different demographic groups in different types of advertising. They were particularly interested in looking at gender differences across advertising related to science as opposed to other topics. They primarily intended to investigate if the interests (or lack thereof) in science advertising would echo societal biases leading to reduced engagement of women in STEM. While they found a range of differences in demographic groups and interest in different kinds of advertising, their study concluded that there is little evidence suggesting that advertising data can measure the decline in interest in STEM for young women in the USA.
2. *Yang, Hui & Menczer (2022)* studied the impact of inauthentic user accounts such as bots on distorting the political discourse in social media. They did so by collecting tweets associated with a 2018 US midterm election. They proposed a method to identify voters on Twitter and systematically compare their behaviours with different random samples of accounts. This then enabled their analysis looking at the effects of inauthentic accounts. They found that some accounts do indeed flood the public data stream with political content, leading to an over-representation of the activity of these accounts in datasets and reducing the visibility of the voice of others.
3. *Smetanin (2022)* provided new resources enabling sentiment analysis in the Russian language, which had been limited to date due to the lack of resources comparable to other languages such as English. The author sampled tweets from the publicly available Twitter Stream Grab to end up with a collection of more than 13K tweets annotated for sentiment by multiple annotators. Tweets were manually categorised into one of five classes: positive, neutral, negative, speech act, and skip. With this study, the author also released a novel transformer-based sentiment classification model for the Russian language.
4. *Rimjhim & Dandapat (2022)* investigated social media data to understand if it can reveal insights about gender-based violence across different countries. To tackle this, the authors look at correlations between social media content linked to gender-based violence and government statistics across different countries. The authors employ a range of methods to analyse these correlations, including graph-based methods. The study concludes that countries with similar cultures show similarity in social media content about gender-based violence, thus validating the hypothesis that content in social media may reflect societal behaviour when it comes to gender-based violence.

5. *Almerekhi, Kwak & Jansen (2022)* investigated changes in online behaviour of users who publish in multiple communities on Reddit by measuring their toxicity levels. They first automatically labelled a large collection of over 87 million posts as toxic or non-toxic, which they then analysed. The study aimed to identify toxicity changes by a user within the same community, across multiple communities, and over time. The study revealed, among others, that a user's toxic behaviour is highly dependent on the toxicity level of the community they participate in, suggesting that users adapt their behaviour to the norm of the communities in question.
6. *Baghdadi et al. (2022)* created a new Twitter dataset with Arabic tweets, where tweets are labelled as suicidal or not. The objective is to develop a model which can detect suicidal ideation from tweets, as a binary classifier that can identify these problematic cases to act upon early on. After collecting a dataset which was labelled by five different annotators, they experimented with a range of models. Among the models under study, they found that a transformer-based model that used Arabic BERT achieved the best performance. Their study is among the first contributions to study suicidal ideation detection in the Arabic language.
7. *Pratama & Firmansyah (2022)* investigated the engagement of political leaders and health organisations with their audience in social media during the COVID-19 pandemic. To measure the engagement, they focus on the concept of online societal association. They collected a dataset with over 173K tweets posted by political leaders and health organisations, which they analysed. Among the political leaders, they found that the one that engaged the most was the Primer Minister of the United Arab Emirates, whereas among the health organisations the one that prevailed was the Public Health Agency of Canada. The study highlights the importance of effective communication through social media during pandemics.
8. *Baxi, Philip & Mago (2022)* examined how the pandemic was reported in the news as well as differences in public reaction to the pandemic in the West and the East. They used archival data from Facebook posts associated with COVID-19 news and posted by English-language mass media between 2020 and 2022, with a total collection of over 700K posts. They use the Valence Aware Dictionary and sEntiment Reasoner (Vader) to measure the news tone and the sentiment polarity score. Their study concludes that posts about the pandemic were more negative in the West than in the East.
9. *Nguyen & Gokhale (2022)* introduced a classification method to detect anti-government sentiment from tweets. They studied the effectiveness of such classification algorithm during COVID-19 anti-lockdown protests in the USA. The authors collected, annotated and released their own datasets associated with two events, namely Operation Gridlock and anti-lockdown protests in Michigan. By combining a range of text-based and user-based features to perform the classification, the study shows that the proposed method could effectively identify anti-government sentiment in tweets with an accuracy of 85% and an F1 score of 0.82.
10. *Ramponi et al. (2022)* studied hate speech expressed in social media which is motivated by religious beliefs. The study makes a distinctive contribution to the hate speech detection literature which had largely focused on other kinds of hate such as racism,

sexism and misogyny, but had overlooked the problem of religious hate. Authors introduce a novel Twitter dataset containing samples of religious hate in two languages, English and Italian, and perform benchmark classification experiments on these two new datasets. They published the datasets aiming to encourage researchers to conduct further research in this understudied problem.

11. *Shamoi et al. (2022)* studied the sentiment expressed by social media users towards vegan diets. They collected a Twitter dataset discussing vegan diets and used a method based on mutual information to extract sentiment-related information from tweets and derived their sentiment towards veganism. Their study revealed that veganism is becoming increasingly popular in recent years and is currently framed more positively than in previous years. In addition to increased positivity, however, they also found increasing rates of expressions of fear towards vegan diets, which shows a potential sign of increased polarisation between supporters and opponents of veganism.
12. *Obeidat et al. (2022)* developed a novel Twitter dataset for misinformation detection from social media, with a specific focus on COVID-19 misinformation. The authors argue that misinformation can occur in many forms and can have different levels of impact on society, and therefore develop a fine-grained annotation schema that includes 19 different types of misinformation. They carefully curated annotation guidelines to enable high-quality annotation within this annotation schema. Following these guidelines, they labelled a dataset with 6.7K tweets which they make publicly available.
13. *Al-nuwaiser (2022)* conducted a user study through crowdsourcing platforms to investigate the impact of using images with news in social media. They investigated news posted on the Facebook platform, and assessed users' perceptions of images of three different types: data visualisation (directly about risk information), advisory (not containing direct risk information, but instead help on how to lower risk), or clickbait (containing no risk-related information, just generic visuals). The study highlights the importance of using images with news reporting in social media, showing the importance of choosing the right type of images.
14. *Heaton et al. (2023)* made a critical reflection on how techniques for computational linguistics are being used to analyse Twitter discourse with different research objectives in mind. They focus on three different techniques for language analysis, which include topic modelling, sentiment analysis and emotion detection. The authors highlight limitations in how these methods are currently being used and how their outputs are being interpreted, suggesting a more careful research design. Among the challenges encountered, they find that the presence of negation and sarcasm in social media posts jeopardises the performance of the aforementioned techniques, which need further exploration.
15. *Ali, Irfan & Lashari (2023)* performed an analysis of sentiment towards COVID-19 countermeasures taken by the Pakistani government during the years of 2020 and 2021. To enable this analysis, the authors collected datasets from the social media platform Twitter using the Snscrape collection tool, which they subsequently annotated. They implemented and compared four state-of-the-art sentiment classification models,

which included Valence Aware Dictionary and sEntiment Reasoner (VADER), TextBlob, Flair, and Bidirectional Encoder Representations from Transformers (BERT). They found that the transformer-based model, BERT, was most accurate among the tested settings, where a fine-tuned BERT model achieved an overall accuracy of 92%.

16. *Hidayatullah et al. (2023)* focused their research on the language identification task. That is, given a text as input, developing a model that automatically identifies the language used in the input text. Their primary focus is on code-mixed language in the cases of the Indonesian, Javanese and English languages. They collect and annotate a Twitter dataset to conduct experiments on language identification. Through testing a range of different state-of-the-art classification models, they find that a fine-tuned IndoBERTweet model performed best.
17. *Fudholi et al. (2023)* suggested the development of a named entity recognition system to facilitate the analysis of tourism-related information from social media. Indeed, information can be overwhelming in social media if not conveniently organised, and therefore named entity recognition could help organise comments by location, for example. To tackle the problem, they developed a BERT-based named entity recognition system, which achieved competitive performance with an F1 score of 0.80. They also conducted a survey to quantify user satisfaction, which was overall above a score of 4 in a scale from 1 to 5.

CONCLUDING REMARKS

The special issue attracted a large number of submissions, of which 17 were ultimately accepted to be included in the issue. These articles include a wide range of topics within the umbrella of analysing and mining social media data. While there is some diversity in terms of social media platforms and languages studied, we believe that these are among the key dimensions that would need expanding in future research.

When it comes to languages, there is an urgent need to analyse, and develop methods for, data written in different languages. This is crucial to enable generalisability of the methods and insights, and to make sure that our analyses cover all regions of the world, including under-represented areas.

Looking at the social media platforms studied, Twitter has been, until 2023, predominantly used as a data source for social media studies. This has been the case primarily because of its generous access to a free API, which provided access to large datasets. As this has recently changed, where Twitter charges a fee for access to its API, which is not affordable for the majority of the scientific community, there is an urgent need to move on and find alternative platforms to research.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

Arkaitz Zubiaga and Paolo Rosso are Academic Editors for PeerJ Computer Science.

Author Contributions

- Arkaitz Zubiaga analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Paolo Rosso analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is an Editorial.

REFERENCES

Al-nuwaiser WM. 2022. Effect of visual imagery in COVID-19 social media posts on users' perception. *PeerJ Computer Science* **8**:e1153 DOI 10.7717/peerj-cs.1153.

Al Tamime R, Weber I. 2022. Using social media advertisement data to monitor the gender gap in STEM: opportunities and challenges. *PeerJ Computer Science* **8**:e994 DOI 10.7717/peerj-cs.994.

Ali MF, Irfan R, Lashari TA. 2023. Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: a study on governmental preventive measures. *PeerJ Computer Science* **9**:e1220 DOI 10.7717/peerj-cs.1220.

Almerekhi H, Kwak H, Jansen BJ. 2022. Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. *PeerJ Computer Science* **8**:e1059 DOI 10.7717/peerj-cs.1059.

Baghdadi NA, Malki A, Magdy Balaha H, AbdulAzeem Y, Badawy M, Elhosseini M. 2022. An optimized deep learning approach for suicide detection through Arabic tweets. *PeerJ Computer Science* **8**:e1070 DOI 10.7717/peerj-cs.1070.

Baxi MK, Philip J, Mago V. 2022. Resilience of political leaders and healthcare organizations during COVID-19. *PeerJ Computer Science* **8**:e1121 DOI 10.7717/peerj-cs.1121.

Fudholi D, Zahra A, Rani S, Huda S, Paputungan I, Zukhri Z. 2023. BERT-based tourism Named Entity Recognition: making use of social media for travel recommendations. *PeerJ Computer Science* **9**:e1731 DOI 10.7717/peerj-cs.1731.

Heaton D, Clos J, Nichele E, Fischer J. 2023. Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. *PeerJ Computer Science* **9**:e1211 DOI 10.7717/peerj-cs.1211.

Hidayatullah AF, Apong RA, Lai DT, Qazi A. 2023. Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets. *PeerJ Computer Science* **9**:e1312 DOI 10.7717/peerj-cs.1312.

Nguyen H, Gokhale S. 2022. An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests. *PeerJ Computer Science* **8**:e1127 DOI 10.7717/peerj-cs.1127.

Obeidat R, Gharaibeh M, Abdullah M, Alharahsheh Y. 2022. Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations. *PeerJ Computer Science* **8**:e1151 DOI 10.7717/peerj-cs.1151.

Pratama AR, Firmansyah FM. 2022. COVID-19 mass media coverage in English and public reactions: a West-East comparison *via* Facebook posts. *PeerJ Computer Science* **8**:e1111 DOI 10.7717/peerj-cs.1111.

Ramponi A, Testa B, Tonelli S, Jezek E. 2022. Addressing religious hate online: from taxonomy creation to automated detection. *PeerJ Computer Science* **8**:e1128 DOI 10.7717/peerj-cs.1128.

Rimjhim , Dandapat S. 2022. Is gender-based violence a confluence of culture? Empirical evidence from social media. *PeerJ Computer Science* **8**:e1051 DOI 10.7717/peerj-cs.1051.

Shamoi E, Turdybay A, Shamoi P, Akhmetov I, Jaxylykova A, Pak A. 2022. Sentiment analysis of vegan related tweets using mutual information for feature selection. *PeerJ Computer Science* **8**:e1149 DOI 10.7717/peerj-cs.1149.

Smetanin S. 2022. RuSentiTweet: a sentiment analysis dataset of general domain tweets in Russian. *PeerJ Computer Science* **8**:e1039 DOI 10.7717/peerj-cs.1039.

Yang K-C, Hui P-M, Menczer F. 2022. How Twitter data sampling biases U.S. voter behavior characterizations. *PeerJ Computer Science* **8**:e1025 DOI 10.7717/peerj-cs.1025.

PeerJ Publishing

PeerJ is a modern and streamlined publisher, built for the internet age. Our mission is to give researchers the publishing tools and services they want, with a unique and exciting experience. All of our seven journals are Gold Open Access and are widely read and cited, with over 500,000 monthly views and 48,500 content alert subscribers. We have published over 19,000 peer-reviewed articles since 2013.


Prestigious Editorial Board


High-Impact Research


Quality Peer Review


Rapid Publishing


Optimum Discoverability

PeerJ Computer Science

High-quality, developmental peer review, coupled with industry leading customer service and an award-winning submission system, means *PeerJ Computer Science* is the optimal choice for your computer science research.

Impact Factor: 3.8
Citescore: 4.2

Scimago Ranking: 0.638
SNIP: 1.094

PeerJ Computer Science

Using social media advertisement data to monitor the gender gap in STEM: opportunities and challenges

Reham Al Tamime and Ingmar Weber

Hamad Bin Khalifa University (HBKU), Qatar Computing Research Institute, Doha, Qatar

ABSTRACT

Boosting the number of women and girls entering careers involving STEM (Science, Technology, Engineering and Maths) is crucial to achieving gender equality, one of the UN Sustainable Development Goals. Girls and women tend to gravitate away from STEM fields at multiple stages from childhood through mid-career. The leaky pipeline is a metaphor often used to describe the loss of women in STEM and arguably other fields before reaching senior roles. Do interests expressed on social media mirror the leaky pipeline phenomenon? In this article, we collected advertisement data (reach estimates) from Facebook and Instagram disaggregated by US metros, age, gender, and interests related to STEM. We computed the Gender Gap Index (GGI) for each US metro and age group. We found that on Instagram, the GGIs for interest in *Science* decrease as users' age increases, suggesting that relatively there is evidence that that women, compared to men, are losing interest in STEM at older ages. In particular, we find that on Instagram, there are plausible relative trends but implausible absolute levels. Nevertheless, is this enough to conclude that online data available from Instagram mirror the leaky pipeline phenomenon? To scrutinize this, we compared the GGIs for an interest in *Science* with the GGIs for placebo interests unrelated to STEM. We found that the GGIs for placebo interests follow similar age patterns as the GGIs for the interest in *Science* across US metros. Second, we attempted to control for the time spent on the platform by computing a usage intensity gender ratio based on the difference between daily and monthly active users. This analysis showed that the usage intensity gender ratio is higher among teenagers (13–17 years) than other older age groups, suggesting that teenage girls are more engaged on the platform that teenage boys. We hypothesize that usage intensity differences, rather than inherent interest changes, might create the illusion of a leaky pipeline. Despite the previously demonstrated value and huge potential of social media advertisement data to study social phenomena, we conclude that there is little evidence that this novel data source can measure the decline in interest in STEM for young women in the USA.

Subjects Data Science, Social Computing, World Wide Web and Web Science
Keywords Advertisement data, Social media, Gender gap in STEM, Interest in STEM, Instagram Ads data, Facebook Ads data, USA, Activity level, Leaky pipeline

Submitted 7 February 2022
Accepted 6 May 2022
Published 21 June 2022

Corresponding author
Reham Al Tamime,
realtamime@hbku.edu.qa

Academic editor
Arkaitz Zubiaga

Additional Information and Declarations can be found on page 24

DOI **10.7717/peerj-cs.994**

Copyright
2022 Al Tamime and Weber

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Gender equality in Science, Technology, Engineering, and Mathematics (STEM) is vital for the achievement of the internationally agreed development goals, including the 2030

Sustainable Development Goals (SDG). Over the past decades, the global community has made concrete effort in inspiring and engaging women and girls in science. Yet women and girls continue to be excluded from participating fully in science (*United Nations, 2022*). On average, only around 30% of the world's researchers are women. Moreover, less than a third of female students choose to study higher education courses in subjects like math and engineering (*Wood, 2021*).

The under-representation of women in STEM fields can be explained by multiple theories, including the "leaky pipeline" metaphor. The leaky pipeline, proposed in 1983, refers to the process by which initially in childhood, people of different genders show similar proportions of interest in the STEM sector. However, gender stereotypes and biases are internalized as they grow up and reach adolescence. This leads to the loss of women in STEM and arguably other fields before reaching senior roles (*Berryman, 1983*).

Modeling the leaky pipeline remains a challenge to apply nationally or globally. Worldwide surveys and censuses are proportionately expensive and time-consuming (*Kashyap et al., 2020*). Towards filling this gap, our article presents a different methodology to mirror the leaky pipeline phenomenon by examining the potential of using online sources of data, particularly data provided by the Facebook and Instagram Advertising Platforms.

Concretely, Facebook and Instagram provide their advertisers access to the number of users through their targeted advertising platform. Before an ad is launched, and before any cost is incurred, advertisers are provided with an estimate of how many users are likely to match specific criteria, which can include age, gender, sets of interests, specific locations, and much more. A recent work (*Vieira & Vasconcelos, 2021*) used Facebook Advertising data to assess the gender gap in STEM in Brazil. Building on this work, we focus on using Advertising data for analyzing the gender gap in STEM across age groups in the USA. The USA is an interesting case study because the majority of Americans say they use Facebook, while the use of Instagram is especially common among adults under 30 (https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/). The USA is also an interesting scenario because although women make up nearly half of the U.S. workforce, they are only 27% of STEM workers (https://www.census.gov/library/stories/2021/01/women-making-gains-in-stem-occupations-but-still-underrepresented.html).

In this article, we investigate the feasibility of using Facebook and Instagram advertising audience estimates for modeling the leaky pipeline phenomenon in the USA. In particular, we filter the advertising audience estimates by location, age, gender, and interests related to STEM. Then, we compute the gender gap index and compare it across age groups at the US metro level. We perform the analysis at the US metro level because a country such as the USA is very diverse and has different cultural norms in different geographical areas. Some of the metros are more conservative, while others are more liberal. Hence, looking at the USA national level might hide regional variation in gender gaps. Unlike previous work, we perform validity checks in the form of estimates of "placebo" interests and usage intensity gender ratio analysis on Instagram. To the best of our knowledge, Instagram ads audience estimates have not been rigorously tested for

data quality and face validity. As such, we demonstrate the dangers of using social media advertisement data as a "black box", and provide methodological tools for future research of such social media data novel source.

RELATED WORK

Gender gap in STEM

Reducing the gender gap in Science, Technology, Engineering, and Mathematics (STEM) is essential to achieve the sustainable developing goals, including Goal 5 on Gender equality. According to the United Nations, women are still excluded from participating fully despite the efforts that have been made to inspire women to study and work in STEM (*Wood, 2021*). Overall, women make up small percentages of STEM graduate students, tenure-track faculty, and tenured faculty (*Ceci et al., 2014*). Women still pursue STEM degrees at much lower rates than their male peers despite the reversal of the gender gap in educational attainment and near gender parity in math performance (*Legewie & DiPrete, 2014*). In most countries, a smaller percentage of women than men graduate with a degree in STEM (*Card & Payne, 2021*). In addition, women worldwide compromise less than 30% of STEM researchers (*United Nations, 2022*). Girls are less likely than boys to aspire to STEM occupations, even when they have comparable abilities and self-evaluations in math and science (*Mann & DiPrete, 2016*).

Research shows that women tend to gravitate away from STEM careers due to a relative abundance (*i.e.*, surplus) of interests that make non-STEM careers equally or more appealing (*Cardador, Damian & Wiegand, 2021*; *Wang & Degol, 2017*). Along these lines, several studies show that women are more interested in jobs involving people and social interactions such as life sciences or health-related professions. In contrast, men are more interested in jobs involving physical objects and abstract concepts such as physics, mathematics, or engineering (*Legewie & DiPrete, 2014*; *Ceci et al., 2014*; *Han, 2016*; *Wang & Degol, 2017*; *Delaney & Devereux, 2019*). As such, for the past three decades, the percentage of bachelor's degrees earned by women in biological sciences and chemistry has increased. However, the percentage of bachelor's degrees earned by women in physics and engineering has stayed stagnant, and in other fields such as computer science, the percentage of degrees earned by women has decreased (*Wang & Degol, 2017*; *Cowgill et al., 2021*).

There are many reasons for the lower representation of women in STEM. Gender stereotypes, gender inegalitarian attitudes or discrimination against women are the main reasons that push women away from STEM fields (*Soylu & Adams, 2020*). Moreover, a survey of college students suggests that lack of interest, feeling unwelcome, and having concerns about the coursework difficulty are reasons that women differentially avoid STEM fields (*Carrell, Page & West, 2010*). Other research finds that lower performance of girls compared with boys in math and science at school, lack of role models (*e.g.*, female teachers/professors in math and science) and self-confidence (*e.g.*, the competitive environment associated with STEM careers) are factors for the lower representation of women in STEM careers (*Zhou et al., 2017*; *Gomez, Alvarado & Nisperuza, 2020*). Similar research tried to summarise six explanations for women's under-representation in math-

intensive STEM fields in the USA. These are (a) cognitive ability, (b) relative cognitive strengths, (c) occupational interests or preferences, (d) lifestyle values or work-family balance preferences, (e) field-specific ability beliefs, and (f) gender-related stereotypes and biases (*Wang & Degol, 2017*). Other research revealed that the under-representation of certain groups in STEM fields is not only a consequence of lower levels of academic qualifications, but also a consequence of lower levels of interest and inclination toward such fields (*Riegle-Crumb & King, 2010*). The research found that interest in STEM is indeed an essential factor to consider as boys often express higher self-efficacy, more joy in science, and a broader interest in science than do girls (*Stoet & Geary, 2018*).

The under-representation of women in science, technology, engineering, and mathematics (STEM) fields can be explained by several theories, including the "leaky pipeline" metaphor. According to this metaphor, women are more likely than men to leave STEM fields at multiple time points from the beginning of college through academic tenure (*Miller & Wai, 2015*). In childhood, girls and boys tend to show similar proportions of interest in the STEM sector. However, this changes as people reach adolescence as gender stereotypes play a role in decreasing the number of women who are interested in STEM. This decrease happens gradually until it reaches tertiary and doctorate levels of education (*Stoet & Geary, 2018*). In other words, the under-representation of women in STEM, especially in computer science, engineering, and physics begins before college and is attributable to a failure to attract women to enter these fields and stay (*Cheryan et al., 2017*). In addition, females leave STEM fields at higher rate than do men even when females perform as well or better than their male peers on STEM-related tests or projects (*Hill, Corbett & St. Rose, 2010*; *Reinking & Martin, 2018*). Accordingly, the pipeline model sheds light on the importance of increasing the volume of flow of females from grade school to graduate school and preventing "leakage" down the line at all stages (*Xu, 2008*).

Several studies have evaluated and monitored the gender gap in STEM worldwide. For example, *Verdugo-Castro, García-Holgado & Sánchez-Gómez (2019)* reviewed instruments and questionnaires related to the gender gap that exists in the field of education within the STEM sector. The review did not only include questionnaires that dealt with the existing gender gap in the STEM sector but also examined particular matters such as the relationship between the choice of studies to be carried out, the influence of parents, self-confidence in the STEM sector, the gender gap in computer engineering, and stereotypes between men and women in adolescence. Focusing on projects funded by the European Union, *García-Holgado et al. (2019)* analysed several research projects about the gender gap in STEM in different calls across the last 5 years. The analysis acknowledged that most of the projects are focused on STEM or STEM subjects, STEM careers, STEM jobs or entrepreneurship, and different technologies such as robotics or IoT. In addition, *Verdugo-Castro et al. (2020)* designed a questionnaire to study gender stereotypes about STEM studies among university students. The study revealed that elements that influence students' opinions are not only the contextual, family, and social ones. Factors also include educational experience and the level of education attained, prior interest in the STEM sector, previous vocational training, and branch of study chosen are all factors that shape the student's opinion of higher STEM studies. *Verdugo-Castro,*

García-Holgado & Sánchez-Gómez (2020) conducted interviews with Spanish women working in STEM. Some of the interviewees admitted that they were encouraged to take STEM and try it out, while others admitted that they were afraid to pursue STEM studies. Similarly, García-Holgado, Díaz & García-Peñalvo (2019) interviewed women as a part of the W-STEM project, which seeks to improve strategies and mechanisms for attracting, accessing, and guiding women in Latin America in STEM higher education programs. These interviews were recorded and shared *via* different channels to make the profiles of these STEM role models visible.

Researchers have offered guidelines and suggestions to attract women to study and work in STEM fields. In particular, Tobar Subía Contento & Nohemi Gamez Aparicio (2020) recommended holding attraction campaigns where the female role models in STEM careers are highlighted, as well as providing data on the number of women studying STEM. Furthermore, the study recommended organizing talks for students to promote vocations in STEM areas. Other studies such as Bennaceur et al. (2018) proposed creating programs and science festivals that encourage girls to study engineering from an early age. The study also proposed developing mentoring programs and promoting the visibility of women as role models through diverse social media channels such as YouTube, Facebook, and Twitter. González-González & García-Holgado (2021) agreed that it is essential to start STEM programs with students at a young age and show them that girls and boys can study and work in STEM. To promote role models in the domain of computer science and software engineering, Ribaupierre et al. (2018) advocated in favour of having a good ratio of women and men on the university teaching team and in favour of good ratio of women and men by software engineering and computer science sub-area. Other work such as Torres-Ramos et al. (2020) stressed the importance of "self-efficacy," known as the ability to accomplish a given task to attract more women and girls in STEM fields. Encouraging and supporting girls' self-efficacy can be achieved by "academic advising, faculty mentorship, tutoring, internship opportunities, and career and skill development" (ibid.). After finding that girls reported significantly lower levels of technical self-efficacy and lower interest in computer science than boys, Brauner et al. (2018) offered several guidelines to adjust children's mental model, including inviting guests from STEM disciplines into the classrooms, organizing Girls' Days, inviting boys and girls into universities for research internships and launching events that help boys and girls understand computing principles and increase their technical self-efficacy.

Quantitative and qualitative techniques have been implemented previously to address the gender gap in STEM. For example, a survey has been used to explore patterns of individuals talented in STEM and determine whether these patterns and experiences differed for men and women or women from different age groups (Heilbronner, 2013). Furthermore, a survey research method was followed to assess the situation of women in STEM in Latin America and Europe and observe if there is any significant gender gap (García-Holgado et al., 2020). Focusing on primary school students, a questionnaire has been designed to study if there are gender differences in aspects related to the experience with mathematics (Ayuso et al., 2021). Focusing on undergraduate students, a survey has been distributed to look at the preponderance of female and male students in science-

oriented majors leading to STEM careers connected to internal and institutional factors (*Faitar & Faitar, 2013*). In addition, a survey was carried out to understand the perception of students/teachers and analyze the factors of the gender gap in STEM (*Moreno et al., 2014*). Other studies relied on existing survey data to study topics related to the gender gap in STEM. For instance, the Freshman Survey data collected by the Cooperative Institutional Research Program (CIRP) was used to analyse how the characteristics of women planning to majors vary across different STEM subfields (*Sax & Newhouse, 2018*). The Engineers Statistical Data System (SESTAT) was also used to assess the factors influencing the gender gap in persistence in STEM employment in computer science and engineering occupations (*Sassler, Michelmore & Smith, 2017*). The Educational Longitudinal Surveys of 2002 to 2012 (ELS) was used to assess the gender gap in STEM major completion (*Weeden, Gelbgiser & Morgan, 2020*), while the 2006 PISA dataset was used to study the gender gap in STEM career expectations across countries (*McDaniel, 2016*). Other researchers used the NCES (National Center for Education Statistics) longitudinal surveys to consider whether males see their life goals as more compatible with STEM careers than females (*Mann & DiPrete, 2013*). Other studies applied pre-post surveys to evaluate different kinds of programs that aim to increase females' interest in and enthusiasm for science through (*Levine et al., 2015*; *Levine & DiScenza, 2018*). Other researchers conducted interviews with women in STEM professions to capture the low participation of women in STEM professions along with the challenges they face in their working environment (*Tandrayen-Ragoobur & Gokulsing, 2021*).

Facebook Ads data

Before showing Ads to Facebook users, Facebook allows advertisers to estimate the potential audience estimate, *i.e.*, estimates of how many people an ad could potentially reach, given specific criteria and targeting options. For example, the monthly potential reach estimate for a Facebook-proper audience, not counting Instagram or Messenger, located in the USA, aged 13 to 50+, and female is 67,000,000. The potential reach audience estimate is available for free in the Facebook Ads Manager (https://www.facebook.com/business/help/200000840044554?id=802745156580214). The estimated audience size can be filtered by age, gender, location, and device type. In addition, the estimated audience size can be filtered by users' interests. The interests filter depends on what Facebook users engage with and like on the platform. Facebook determines a user's interests based on the pages they like, the content they view, and the ads they click (https://www.facebook.com/help/562973647153813).

Facebook's audience estimates have been used for a range of studies connecting society and the world's online population. For example, researchers took advantage of Facebook's audience estimate to predict gender gap in internet use (*Kashyap et al., 2020*). The monthly active Facebook users have been filtered by age, gender, and country. The study demonstrated that advertising digital data from Facebook can be used to complement traditional data sources to monitor global development indicators linked to digital gender inequality. Similar research worked on filtering Facebook Ads data by gender, geography, age, education, industries, and device type in order to examine

variation in gender gaps in India (*Mejova et al., 2018*). The study found that there is a significant sub-national variation in gender gaps in Facebook use and higher gender gaps tend to be present in states with greater gender inequality in schooling and lower levels of social and economic development. Other researchers developed a metric called Facebook Gender Divide (FGD) to measure gender differences in Facebook access and activity in 217 countries (*Garcia et al., 2018*). The FGD tends to be associated with other types of gender inequality, including economic, health, and education inequality. Recently, researchers have looked into using Facebook Ads data to measure the gender gap in mobile device usage duration and saw patterns of gender gaps disfavouring women, especially among younger people (*Sabri, Kashyap & Weber, 2021*). In a different article, *Gil-Clavel & Zagheni (2019)* extended the scope and used Facebook Ads data to include the role of age. The authors described that in countries in North America and Northern Europe, patterns of Facebook adoption do not differ significantly between older and younger adults. In Asian countries, which have high levels of gender inequality, differences in levels of Facebook adoption by gender disappeared at older ages (ibid.).

Facebook Ads data has also been leveraged to study culture and assimilation. In a recent study, Facebook data was filtered by geography and interests related to specific cultural traits, providing a collection of massive data on human behaviors that are used to complement traditional cultural metrics (*Obradovich et al., 2020*). Moreover, Facebook Ads data has been filtered by geography and interest in Brazilian dishes in order to focus on how a foreign culture such as the Brazilian culture is spread around the world (*Vieira et al., 2020*). In parallel, researchers found that Facebook Ads data can be used to explore the cultural assimilation of Mexican Immigrants by looking at musical preferences. Accordingly, Facebook Ads data has been filtered by interest in musical genres and Facebook's ethnic affinities demographic attributes (*Stewart et al., 2019*). Also, researchers disclosed that Facebook data can provide insights into the assimilation of Arabic-speaking migrants in Germany. For this, Facebook data has been collected and filtered by interests and by different populations (*e.g.*, non-expats living in Germany, or Arabic-speaking expats living in Germany) (*Dubois et al., 2018*).

Facebook Ads data has been used to reinforce research in the political science and health science domains. For instance, the research looked into leveraging Facebook Ads data to infer the audience demographics of politicians in the electoral race (*Ribeiro, Kansaon & Benevenuto, 2019*). Here, Facebook Ads data has been filtered by interests related to candidates to calculate audience demographics. In the context of health, Facebook Ads data has been collected to compute the proportion of users in a target demographic group with interest related to schizophrenia (*Saha et al., 2017*). This data has been used to construct a plausible index of population-scale schizophrenia awareness. Furthermore, Facebook Ads data has been utilized to track health conditions associated with tobacco use, obesity, and diabetes (*Araujo et al., 2017*). This research contributes to previous work by performing quality check in the form of comparing the performance of placebo interest. In consideration of that, Facebook Ads data has been collected and filtered by country, interests in health conditions, interest in placebo, age, and gender. Equivalently, previous work introduced the use of placebo interests and compared these interests to interests

related to four diverse health conditions: (1) diabetes (type II), (2) obesity, (3) food sensitivities, and (4) alcoholism (*Mejova, Weber & Fernandez-Luque, 2018*). The study showed that placebo interest such as Reading and Technology display a nontrivial correlation with other health interest such as Diabetes and Obesity. This makes measuring placebo interests a critical step in verifying the significance of health-specific results.

Assessing the gender gap in STEM has been scrutinized using Facebook Ads data. Focusing on Brazil as a case study, *Vieira & Vasconcelos (2021)* argued that even with a larger proportion of women on Facebook, STEM interests are still concentrated towards men. In particular, Facebook Ads data disclosed that college majors related to Environmental Science, Engineering, and Computer Science have more men interested in them. However, college majors related to Life Science and Math/Physical are preferred by women (ibid.). Additionally, *Vieira & Vasconcelos (2021)* (ibid.) applied gender balance analysis on demographic subgroups, in particular for education levels and age groups. The analysis confirmed that the gender gap in STEM widens significantly in the transition from Bachelor's to post-graduate levels (*e.g.*, Masters or Doctorate levels) as well as into research careers. The authors also compared Facebook users' age with their college major interests. Explicitly, the authors split the population into four subgroups: adolescent (13–19 years), early adulthood (20–39 years), adulthood (40–64 years), and maturity (65 years or more). The authors observed the same pattern as the educational level where the proportion of female Facebook users interested in each major decreases as older as they are. Over their lifetime, women tend to lose even more interest in STEM majors such as Aviation and Mechanical Engineering.

LinkedIn ads data

Similar to Facebook, LinkedIn allows advertisers to estimate the reach of their campaigns. For instance, in 2017 LinkedIn estimated that an advertisement "targeted to 18- to 24-year-old males in San Francisco Bay Area with knowledge in Java, has the potential to reach 11,000 people" (*Haranko et al., 2018*). Such estimates of employment and skills are hard to obtain using other traditional data sources. LinkedIn advertisement data has been filtered by skills, industry, age, location, and gender in order to understand gender gaps in employment along various dimensions (ibid.). The study reported that Education and and Medical/Health care seem to be the most female dominant categories, whereas Construction and Manufacturing seems to be the most male dominant categories. The study also reported that there is a high correlation between LinkedIn data and traditional data source provided by the US Bureau of Labor Statistics. Building on this work, LinkedIn data has been disaggregated by job seniority, job function, field of study, industry, age, location, and gender to examine gender inequalities in the professional domain (*Kashyap & Verkroost, 2021*). This research outlined that gender inequality on LinkedIn is exceptionally high among older individuals and in Africa, the Middle East, and parts of Asia. The study agrees that LinkedIn data is correlated with traditional data sources available from the International Labour Organization's (ILO) Statistical Database. Furthermore, LinkedIn data has been disaggregated by industry, location, and gender to examine variation in the professional gender gap in IT industry (*Verkroost et al., 2020*).

The authors observed that in the large majority in countries worldwide, there are more men than women in IT. Specifically, there are almost four times more men than women in IT industries like networking and computer hardware, while there are almost twice as many men as women in the telecommunication and internet industries. The authors also observed that LinkedIn data tend to be similar to other traditional data sources available from Eurostat, at least for European countries.

The previous studies discussed above use LinkedIn and Facebook Ads data to monitor digital and employment gender gaps, culture, politics, health, and wellbeing. Our study is instead focused on the combination of Facebook and Instagram data for measuring gender gap in STEM. Different from research that explores the gender gap in STEM in Brazil, this research examines the gender gap in STEM in the USA. Specifically, our study is different from *Vieira & Vasconcelos (2021)*'s work in that it considers different geography, platforms, and age groups. We aim to focus on the gender gap in STEM among teenagers and young adults. Our work builds on previous research strands by comparing gender gap in advertisement data across age and USA metropolitan area. Additionally, our research introduces measuring placebo interests and usage intensity gender ratio in order to validate findings related to gender gap in STEM. This is an important methodological contribution as it shows that many apparent differences in terms of interest profiles might be explained by differences in time-spent-on-platform.

METHODOLOGY

Data collection

Facebook and Instagram data has been retrieved from Facebook's Marketing application programming interface (API). The Web interface of the Ads Manager (https://www.facebook.com/business/tools/ads-manager) can render equivalent data, but using the API makes programmatic access easier (*Mejova, Weber & Fernandez-Luque, 2018*). Figure 1 shows a screenshot of the Facebook Ads Manager illustrating the potential reach to an audience likely to match targeting criteria. The Facebook marketing API provides two metrics: Daily Active Users (DAUs) and Monthly Active Users (MAUs). On Facebook for developers, DAUs are defined as the "estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past day" (https://developers.facebook.com/docs/marketing-api/reference/ad-campaign-delivery-estimate/), while MAUs are defined as the "estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past month" (https://developers.facebook.com/docs/marketing-api/reference/ad-campaign-delivery-estimate/).

The DAUs and MAUs reach estimates have been queried from the API and filtered by location, age, gender, interest, and platform as the following:

Location: This covers all the USA metropolitan statistical areas (https://en.wikipedia.org/wiki/Metropolitan_statistical_area) (210 metros). We collected geographical data to capture variations in gender gaps in the USA.

Age: This covers various age groups (13–16, 14–17, 15–18, …26–29 years old) as well as all age groups (13–29 years old).

Gender: This covers female, male, and all genders.

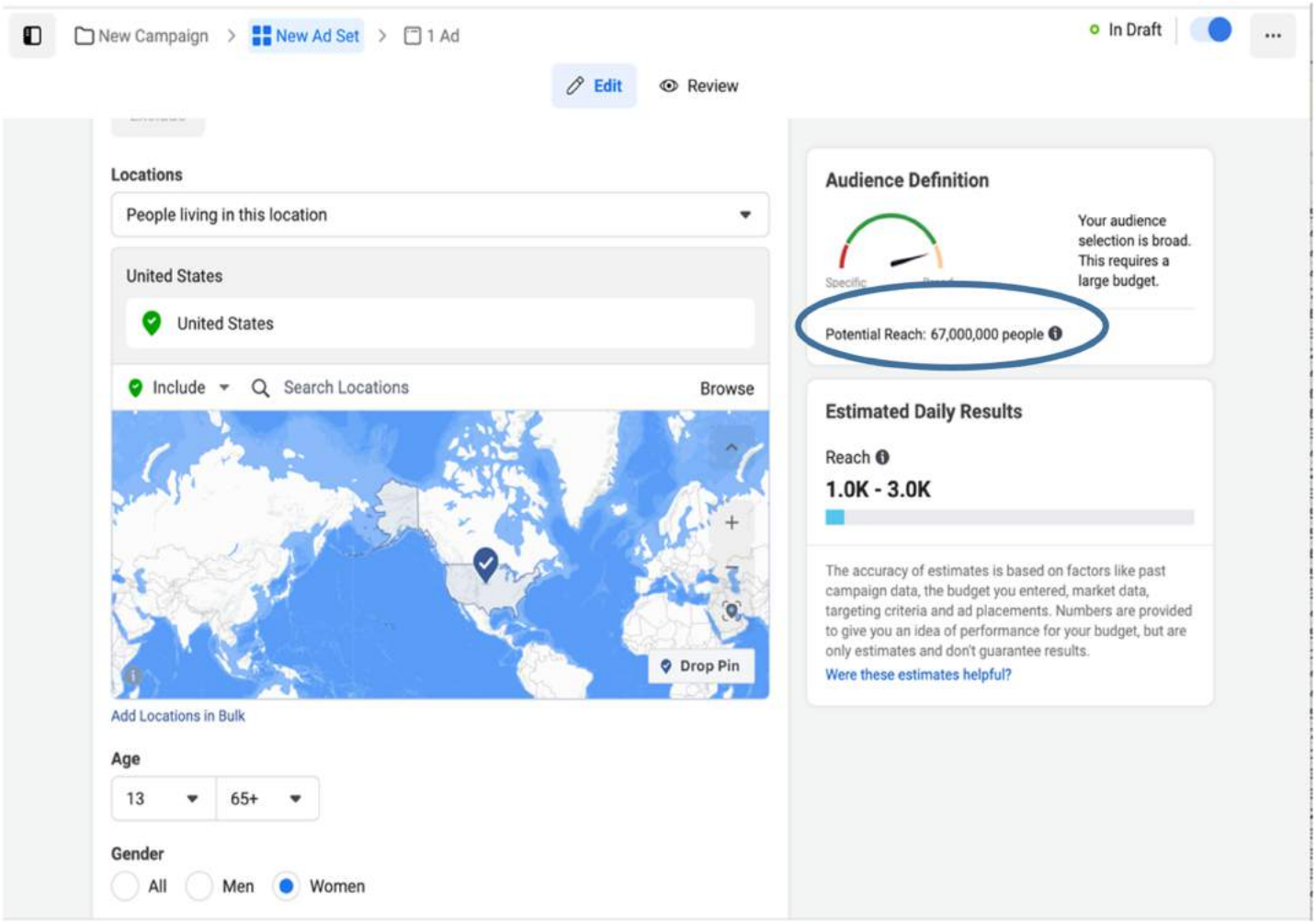


Figure 1 **Facebook Ads Manager.** Full-size DOI: 10.7717/peerj-cs.994/fig-1

Interest: We used the Facebook Ads Manager interface to identify the key interests related to STEM. As such, we filtered the reach estimate by four interests: *Science, Technology, Engineering*, and *Mathematics*.

Platform: This covers two platforms: Facebook (this includes all platforms such as Facebook, Instagram, Audience Network and Messenger), and Instagram-only.

We included Instagram-only as Instagram is known to be popular with teenagers (https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/.). The MAUs metric does not report numbers under 1,000 to prevent the targeting of small groups of individuals. We are including MAUs values under 1,000 in the analysis, but we are replacing these small values with NAs (On Facebook, we got 3,881 NAs out of 12,600 data points. On Instagram, we got 4,966 NAs out of 12,600 data points).

We also collected the number of users not filtered by any interest from Facebook and Instagram, in order to calculate the Gender Gap index.

Table 1 Top and bottom three GGI values for age group 13–16 on Facebook and Instagram (interest in Science).

Platform	Top metros	Bottom metros
Facebook	Birmingham (Ann and Tusc) (1.4)	Harlingen-Wslco-Brnsvl-Mca (0.982)
	Miami-Ft. Lauderdale (1.438)	Fresno-Visalia (1.006)
	Orlando-Daytona Bch-Melbrn (1.516)	San Diego (1.01)
Instagram	Chicago (1.611)	Salt Lake City (1.095)
	Seattle-Tacoma (1.626)	Minneapolis-St. Paul (1.307)
	Dallas-Ft. Worth (1.642)	San Francisco-Oak-San Jose (1.339)

Gender gap index

We have used the MAUs to calculate the Gender Gap Index (GGI) for each USA metro and age group as the following:

$$\text{GGI} = \frac{\text{Female to male gender ratio filtered by each STEM interest}}{\text{Female to male gender ratio of the number of users in each platform}} \quad (1)$$

A GGI less than 1.0 indicates a gender gap with women at a disadvantage. Conversely, a value equal to or larger than 1.0 indicates that gender parity for the interests considered has been achieved or exceeded.

RESULTS

Min and max GGIs

In this section, we report the min (= worst for women) and max (= best for women) GGI values for age groups 13–16 and 26–29 years on Facebook and Instagram (interest in *Science*). On Facebook, the min GGI for age group 13–16 years belongs to Harlingen-Wslco-Brnsvl-Mca (0.982), while the max GGI for the same age group belongs to Orlando-Daytona Bch-Melbrn (1.516). For age group 26–29 years, the min GGI belongs to Laredo (0.885), while the max GGI for the same age group belongs to Greenwood-Greenville (1.49). On Instagram, the min GGI for age group 13–16 years belongs to Salt Lake City (1.095), while the max GGI for the same age group belongs to Dallas-Fort Worth (1.642). For age group 26–29 years, the min GGI belongs to El Paso-Las Cruces (1), while the max GGI for the same age group belongs to Columbus-Tupelo-West Point-Houston (1.725). Table 1 shows the top and bottom three GGI values for age group 13–16 years old, while Table 2 shows the top and bottom three GGI Values for age group 26–29 years old. We expected to see Metros with low GGI (<1), especially with small Metros that might have higher gender gaps. However, this is not the case as there are many GGI values that are >1 in Facebook as well as Instagram (On Facebook, we obtained 4,092 GGI values that are <1, while we obtained 4,627 GGI values that are ≥1 out of 12,600 data points. On Instagram, we obtained 2,862 GGI values that are <1, while we obtained 4,772 GGI values that are ≥1 out of 12,600 data points. These GGIs include interests related to *Science, Technology, Engineering*, and *Mathematics*).

Table 2 Top and bottom three GGI values for age group 26–29 on Facebook and Instagram (interest in science).

Platform	Top metros	Bottom metros
Facebook	Duluth-Superior (1.467)	Laredo (0.885)
	Columbus-Tupelo-W Pnt-Hstn (1.477)	San Angelo (0.936)
	Greenwood-Greenville (1.49)	El Paso (Las Cruces) (0.939)
Instagram	Columbia-Jefferson City (1.608)	El Paso (Las Cruces) (1)
	Monroe-El Dorado (1.701)	Monterey-Salinas (1)
	Columbus-Tupelo-W Pnt-Hstn (1.725)	Utica (1.019)

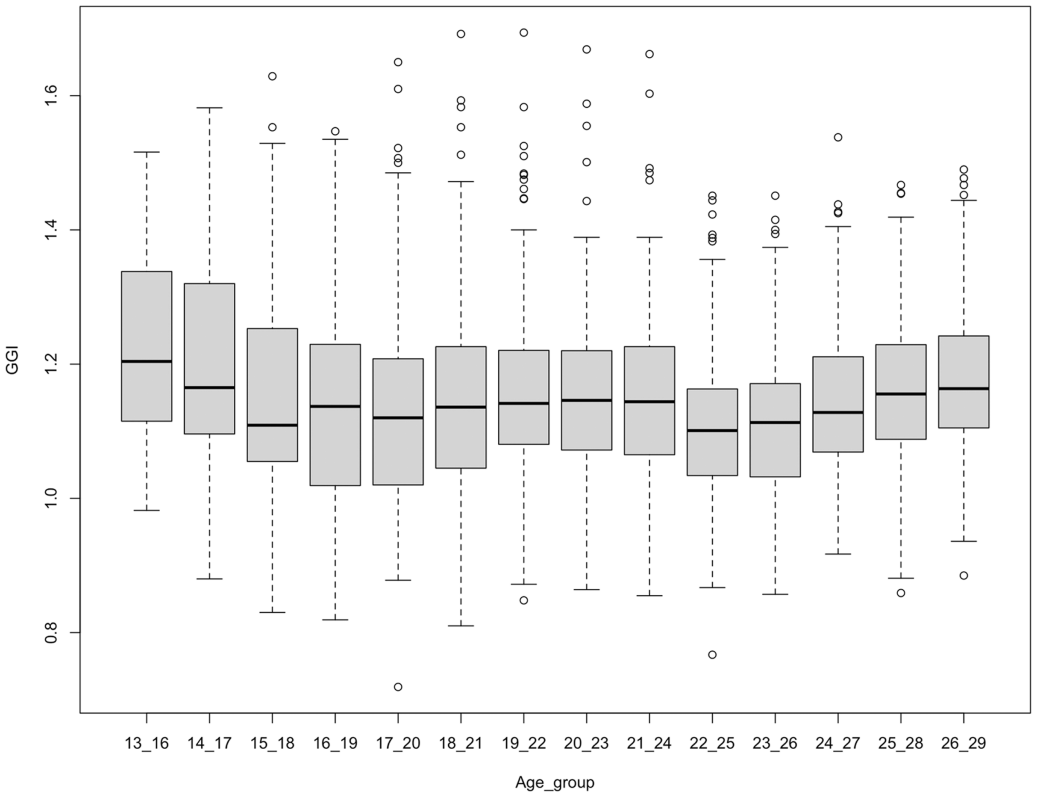


Figure 2 GGI for interest in science across age groups on Facebook. Each point corresponds to a different US metro. Full-size DOI: 10.7717/peerj-cs.994/fig-2

Comparison between GGIs across age groups

Here, we compare the GGI across age groups on each platform. On Facebook, we found that the GGI of interest related to *Science* fluctuates across age groups (Fig. 2). Each point on the boxplot corresponds to a different USA metro. The GGI decreases from age group 13–16 to age group 15–18, then increases from age group 16–19 to age group 21–24. The GGI decreases again at age group 22–25 and then increases at age group 23–26 until reaching age group 26–29 years. We have conducted two tests: the Kruskal-Wallis and the Paired t-test. The Kruskal-Wallis is conducted in order to

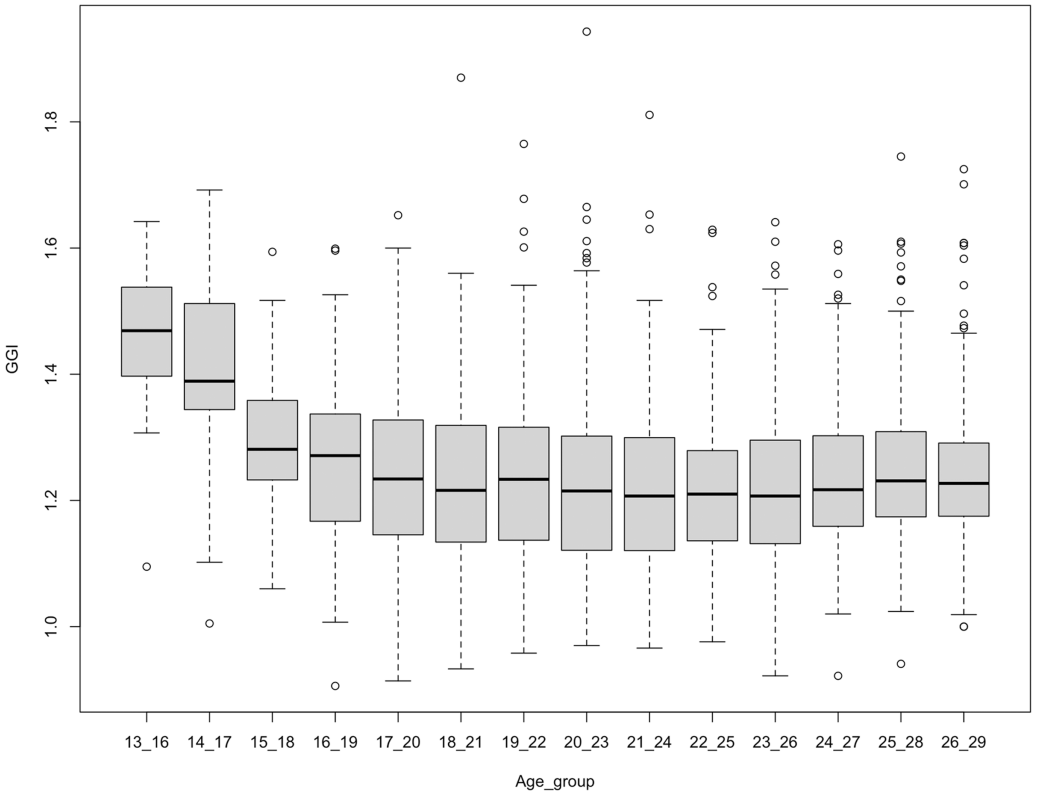


Figure 3 GGI for interest in science across age groups on Instagram. Each point corresponds to a different US metro. Full-size DOI: 10.7717/peerj-cs.994/fig-3

determine if there is a statistically significant difference between the GGIs for different age groups. The paired t-test is conducted in order to determine if there is a difference between the GGIs at age groups 13–16 and 26–29. After conducting Kruskal-Wallis on Facebook data, we reject the null hypothesis and conclude that not all the group medians are equal (Kruskal-Wallis chi-squared = 85.425, df = 13, *p*-value = 1.041e−12, *n* = 2,940, effsiz = 0.0248). The paired t-test results designates that there is significant evidence of change in GGIs between age groups 13–16 and 26–29 years old (t = −5.0328, df = 35, *p*value = 1.453e−05, 95% CI [−0.15179885 to −0.06453449], mean of the differences = −0.108, *n* = 210, effsiz = 0.839). The mean of the difference in GGI between the two age groups is reduced by 0.11 points. Concerning other interests on Facebook, such as interest in *Technology, Engineering*, and *Mathematics*, we did not observe a decrease in the GGI as users get older (Fig. A1). On Instagram, however, we noticed that the GGI of interest related to *Science* is decreasing across age groups, suggesting that women, compared to men, are relatively losing interest in STEM at older ages (Fig. 3). There was a statistically significant differences between GGI in different age groups as assessed using the Kruskal-Wallis test (Kruskal-Wallis chi-squared = 127.47, df = 13, *p*-value = 6.67368e−21, *n* = 2,940, effsiz = 0.0391). The carried out Paired t-test reveals that the mean paired difference between between age groups 13–16 and 26–29 years does not equal zero (t = −8.2268, df = 17, *p*value = 2.489e−07, 95% CI [−0.3296102 to −0.1950564], mean

of the differences = −0.262, n = 210, effsiz = 1.94). The GGI between the two age groups is reduced by 0.26 points. Concerning interests on Instagram related to *Technology, Engineering*, and *Mathematics*, we did not notice a significant decrease in GGI across ages, as shown in (Fig. A2). Overall, the results denote that the GGIs on Instagram for interest in *Science* only mirror the leaky pipeline as women are losing interest in *Science* as they grow up. Specifically, (i) the relative trend across the ages is in line with a leaky pipeline, but (ii) the absolute level (actual GGI values remains > 1.0 all the time) is not in line with a leaky pipeline. Before confirming this outcome and trying to understand the trend on the relative level, we collect and analyse data related to *placebo* interests on Instagram.

Placebo interests

Figure 3 provides potential evidence for a leaky pipeline on Instagram as the GGI is decreasing for older age groups, indicating a relative loss in *Science* of women compared to men. However, to rule out certain confounding factors, we also collected data and performed an analysis for a set of placebo interests. Placebo interests are interests on Instagram that should not have an obvious causal link with interests in STEM but that might still turn out to be correlated due to latent factors such as general engagement on the platform (*Mejova, Weber & Fernandez-Luque, 2018*). Placebo interests are helpful to understand if we are really measuring interest in science on Instagram and not the level of activity on the platform or other unknown latent factors. Intuitively, "these interests are meant as a placebo wherein no topic-specific treatment is performed, and any effect observed is due to the random or causal factors outside the topic" (ibid.). Similar to previous work (ibid.), we used popular generic interests that are not *a priori* linked to the four STEM interest studied on Instagram. The generic interests include *Facebook, Reading, Music*, and *Entertainment*. We also added other gendered interests such as *Cars* and *Beauty*. As many of these Placebo interests are very general with audiences of millions of users worldwide, they can be seen as a kind of proxy for general engagement on the platform: if a user really shows no sign of interest in, say, any type of music, then that could be plausible due to a lack of engagement in activities on the platform. We collected the reach estimates from the API filtered by location, age, gender, and placebo interest as the following:

Location: This covers all the USA metropolitan statistical areas (210 metros).

Age: This covers various age groups (13–16, 14–17, 15–18, …26–29 years old).

Gender: This covers female, male, and all genders.

Interest: We used the Facebook Ads Manager interface to exhaustively enumerate Placebo interests on Instagram. Then, we filtered the reach estimates by interest in *Facebook, Reading, Music, Entertainment, Cars* and, *Beauty*.

The MAUs have been used to calculate the gender gap index (GGI) for each USA Metro and age group as the following:

$$\text{Placebo GGI} = \frac{\text{Female to male gender ratio filtered by each placebo interest}}{\text{Female to male gender ratio of the number of users in each platform}} \quad (2)$$

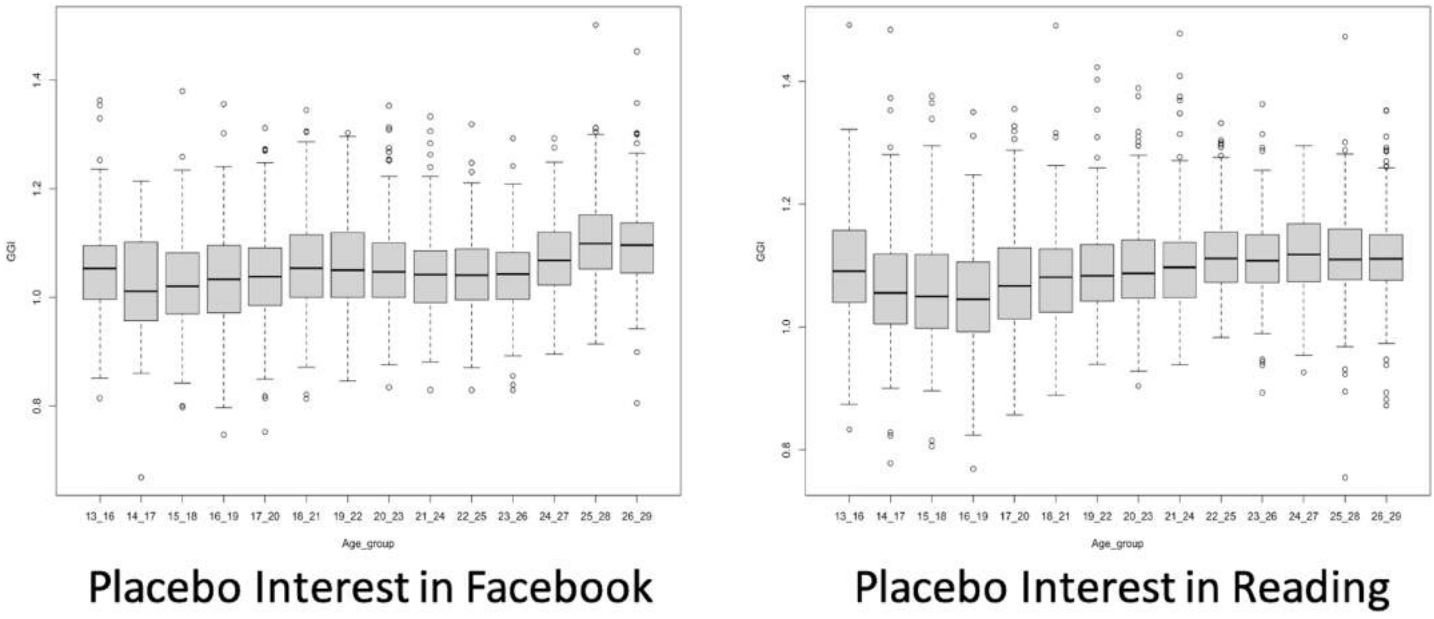


Figure 4 Placebo GGIs for interests in 'Facebook' and 'Reading'. Each point corresponds to a different US metro.
Full-size DOI: 10.7717/peerj-cs.994/fig-4

We discern that the placebo GGIs for interest related to *Facebook* and *Reading* increase as users get older, demonstrating that women, compared to men, are not losing interest in *Facebook* and *Reading* at older ages (Fig. 4). Nevertheless, we discern the placebo GGIs for interest related to *Music, Entertainment, Cars* and, *Beauty* decrease as users get older, showing that women, compared to men, are losing interest in *Music, Entertainment, Cars* and *Beauty* at older ages (Fig. 5). This makes the relative GGIs patterns that we observed on Instagram for interest in *Science* not unique as placebo interest follow similar patterns across age groups. For example, Fig. A3 shows that the GGIs for interest in *Science* for different age groups on Instagram have a similar pattern as the GGIs for interest in *Music*. In fact, we also observed that the 26–29 age group GGI for interest in *Science* is correlated with the GGI for interest in Facebook ($r = 0.46$), interest in Reading ($r = 0.41$), and interest in Music ($r = 0.40$). In addition, we observed that the 13–29 age group GGI for interest in *Science* is correlated with the GGI for interest in Facebook ($r = 0.33$), interest in Reading ($r = 0.41$), and interest in Music ($r = 0.23$). Within placebo interests, we observed that the 13–29 age group GGI for interest in Facebook is correlated with the GGI for interest in Reading ($r = 0.66$), interest in Entertainment ($r = 0.63$), and interest in Music ($r = 0.63$). In general, as shown in Fig. A4, the GGIs for interest in STEM (with the exception of interest in *Math*) are correlated with the GGIs for interest in placebo. As such, we cannot conclude that online data provided by Instagram reflect the leaky pipeline phenomenon when we measure GGI for interest on *Science*.

Activity level on the platform

In the previous section, we found that the GGIs for age groups 13–16 and 14–17 years are higher than the GGIs for the other older age groups when we consider interests in

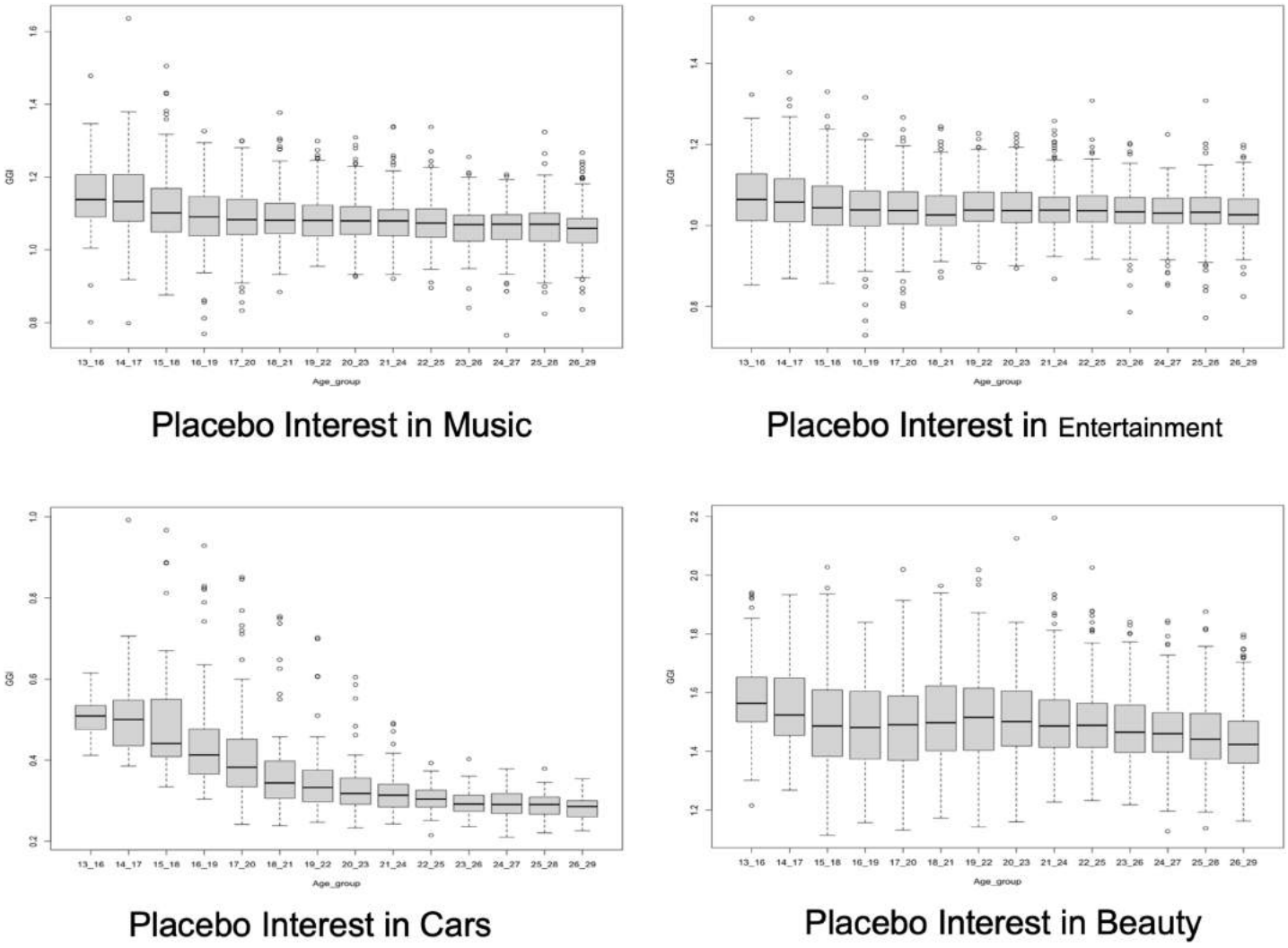


Figure 5 Placebo GGIs for interests in 'Music', 'Entertainment', 'Cars' and 'Beauty'. Each point corresponds to a different US metro.
Full-size DOI: 10.7717/peerj-cs.994/fig-5

Science as well as interests in placebo on Instagram. This difference in the GGIs could be attributed to 13–17 years old users spending more time on the platforms. Unfortunately, time spent on the platforms or usage intensity is not given by Facebook or Instagram by default. So in an attempt to control the time spent on the platform, we look at the DAUs to MAUs ratio on Instagram as:

$$\text{Usage Intensity Gender Ratio} = \frac{\text{DAUs female to male gender ratio of the number of users on Instagram}}{\text{MAUs female to male gender ratio of the number of users on Instagram}} \quad (3)$$

In other words, DAUs to MAUs ratio serves as a proxy for the Usage Intensity Gender Ratio. Equation (3) captures female usage intensity divided by male usage

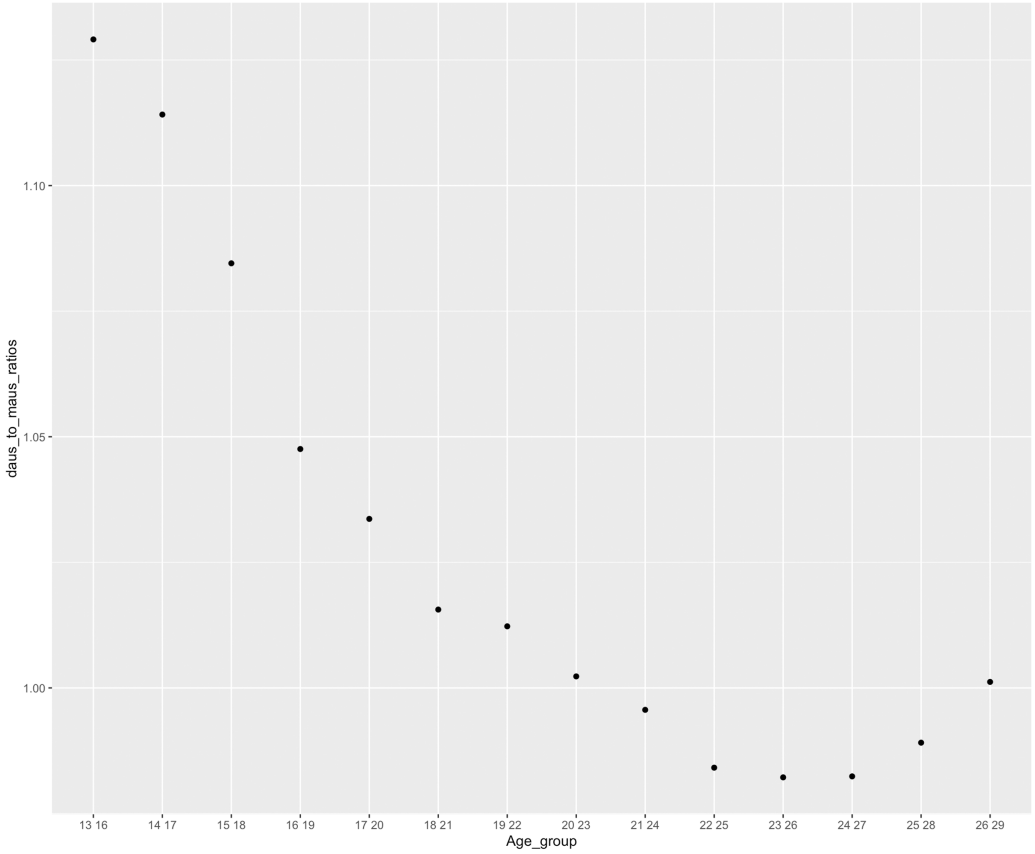


Figure 6 **Usage intensity gender ratio across age groups (US national level).**
Full-size DOI: 10.7717/peerj-cs.994/fig-6

intensity. For user groups who spend more time on the platform, we would expect a bigger proportion of the monthly active users (MAUs) also to be daily active users (DAUs). Correspondingly, we expect a higher DAUs to MAUs ratio and thus higher usage intensity gender ratio. As indicated in Fig. 6, age groups 13–16 and 14–17 years have higher DAUs to MAUs ratios than the other age groups thus, girls who are 13–17 years old are more likely to spend more time on the platform. The more time that users spent on the platform, the more Instagram or Facebook assigns interest for users. For example, if we have two users (User A and User B) within the same age group, but User A spends more time on Instagram than User B. Then, Instagram is more likely to assign more interests to user A. These assigned interests are possibly false positive and do not accurately represent users' actual interests. Facebook and Instagram use machine learning algorithms to infer interests about its users (https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads). Users who spend more time on the platform are also more expected to engage more on the platforms (*i.e.*, like and comment on posts). The more interactions on the platform, the more likely the algorithm is run again and again to assign interest to users. For instance,

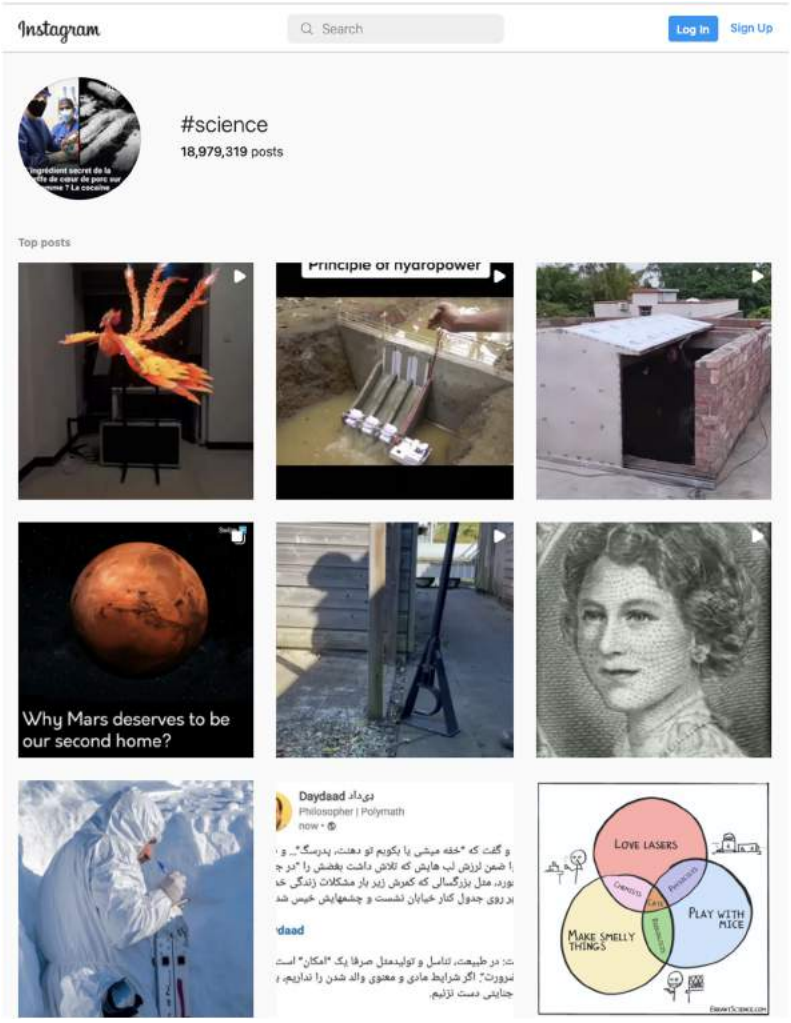


Figure 7 **The hashtag #science on Instagram.** Full-size DOI: 10.7717/peerj-cs.994/fig-7

users who are 13 years old are more engaged on the platform, so Instagram picks more interest for them related to both STEM and Placebo. These teen users are expected to be targeted based on interests that do not reflect their actual interests. This is why we saw that for interests in *Science* and interests in placebo, the GGIs for age groups 13–16 and 14–17 years are higher than the GGIs for the other older age groups. In all likelihood, this stipulates what we are measuring on Instagram is time spent on the platform and usage intensity and, not actually interest related to STEM.

Science hashtag on instagram

In order to explore what is classified as *Science* on Instagram, we also looked at pictures under the hashtag #science. Figure 7 shows that there is a variety of things covered under the *Science* hashtag on Instagram. These things could be directly related to science (*e.g.*, pictures illustrating the principles of hydro-power) and unrelated to science (*e.g.*, images

of a door and roof). This raises many questions about how Instagram defines *Science* and how it assigns interests related to science to users. This also suggests that what we are measuring on Instagram is not necessarily related to STEM or the leaky pipeline. We also inspected pictures under the hashtags #technology, #engineering, and #math on Instagram (Fig. A5) to understand what might be classified as being related to an interest in STEM. As with #science, we found that pictures under those hashtags often do not appear to be related to STEM.

DISCUSSION

This article explores the use of Facebook and Instagram Ads audience estimates for modeling the gender-specific decline in interest in STEM across age, *i.e.*, modeling one element of the leaky pipeline phenomenon. At the USA Metro level on Facebook, women, compared to men, are not losing interest in either *Science, Technology, Engineering*, and *Math* at older ages. On Instagram, however, the relative trend indicates that women, compared to men, are losing interest in *Science* at older ages. While this relative trend across the ages is in line with a leaky pipeline, the absolute levels observed are not: even at older ages, the data suggests higher levels of interest for women than for men. In order to understand this more, we assess the quality of Instagram data by introducing placebo baselines, and usage intensity gender ratio analysis. Among our findings, we show that interests in Placebo such as interest in *Music* display similar patterns as interest in *Science* across age groups on Instagram. Moreover, we show that teenagers (13–17) years old have higher usage intensity gender ratio than older users. As teens spend more time on the platform, this makes them vulnerable to being assigned more interests that are attached to their profiles. This could suggest that we are using the marketing APIs as a tool to measure activity level on the platform, and not interest in STEM. We also looked at pictures under the hashtag #science on Instagram. Despite that the hashtag covers a wide range of things that could be classified under *Science*, there are pictures that are not entirely related to Science or STEM in general. This indicates that it is hard to obtain a definition or classification of *Science* from Instagram. While it is known that, at a high level, Facebook and Instagram assign interests based on users' publications and activities on the platform (https://www.linkedin.com/pulse/how-facebook-defines-interests-targeting-irina-yandulina), the details of the Instagram interest classification algorithm remain a black box. Facebook and Instagram do not only track users on their platform, but also track web behavior outside of their own platform. This includes tracking users across other websites and services, into the various apps they are using on their phone and to the places they physically visit in the real world—especially if they decide to check in on Facebook while they are there (https://www.wired.com/story/ways-facebook-tracks-you-limit-it/). For example, *Aguiar et al. (2022)* showed that Facebook can track 55% of the websites visited by Facebook users, and 44% of non-Facebook users, which amounts to 41% and 38% of browsing time, respectively. Therefore, profiling users is far beyond the science hashtag example given in the article. Recently, Apple released an update for iPhones with a new popup that asked users if they wanted to allow apps on their phones to

target the user for ads (https://www.cnbc.com/2021/11/13/apples-privacy-changes-show-the-power-it-holds-over-other-industries.html). This change in Apple's privacy could lead to the suffering of the Facebook Ads targeting. More important, similar patterns that are observed between interests in Science and interest in Placebo or usage intensity warrant extra caution in the use of Marketing APIs as a source of social interest.

Facebook advertisement data that are filtered by interests have shown promise in measuring schizophrenia awareness (*Saha et al., 2017*), investigating culture at the country, subnational, and local levels (*Obradovich et al., 2020*), assessing how the Brazilian culture is distributed around the world (*Vieira et al., 2020*), and exploring the cultural assimilation of Mexican Immigrants (*Stewart et al., 2019*). In spite of that, this study shows that there is no strong evidence to conclude that online advertisement data that are filtered by interests can be used to study the gender gap in STEM in the USA. These interests in STEM do not reflect users' actual interests and are more prone to be false positives. This study contributes to previous work by not only combining Facebook and Instagram advertisement data to look at the gender gap in STEM in the USA, but also evaluating these novel online sources and comparing them to placebo and usage intensity gender ratios.

Still, it may be too early to give up on this potentially rich data source. Our effort is restricted to specific interests in STEM; we acknowledge that it is impossible to generalize the limitations we have found without future studies. Future work, for example, could look at partitioning the data based on the size of the US metro and comparing the gender gap index across age groups accordingly. Furthermore, to assess the accuracy of interest assignment, one could use the advertising platform to run a survey assessing the interests of the reached audience directly. The DAUs and MAUs change over time, with DAUs changing several times within 24 h and MAUs typically changing every 1–2 weeks. This is both a strength, in terms of recency, as well as a limitation, in terms of time-dependency of the analysis. Furthermore, the algorithms used to infer user attributes can occasionally change, which can make it hard to obtain consistent and comparable signals over time. For example, (*Palotti et al., 2020*) observed a sudden change in March 2019 in how Facebook classifies users by countries they have lived in. Unfortunately, Facebook does not provide access to historical data to show changes over time. Therefore, long-term effort is required to monitor the data with respect to seasonal variation across, say, 1 year. Finally, replicating our analysis using placebo interests in other countries could help with understanding whether interest-specific, rather than usage-related, conclusions can be drawn in other contexts.

We hope that this work will encourage future efforts to use our methodology to gather user interest from Facebook and Instagram for other use cases and research problems. We also hope that the results presented here will encourage future researchers to test the face validity of these potentially rich data sources.

APPENDIX

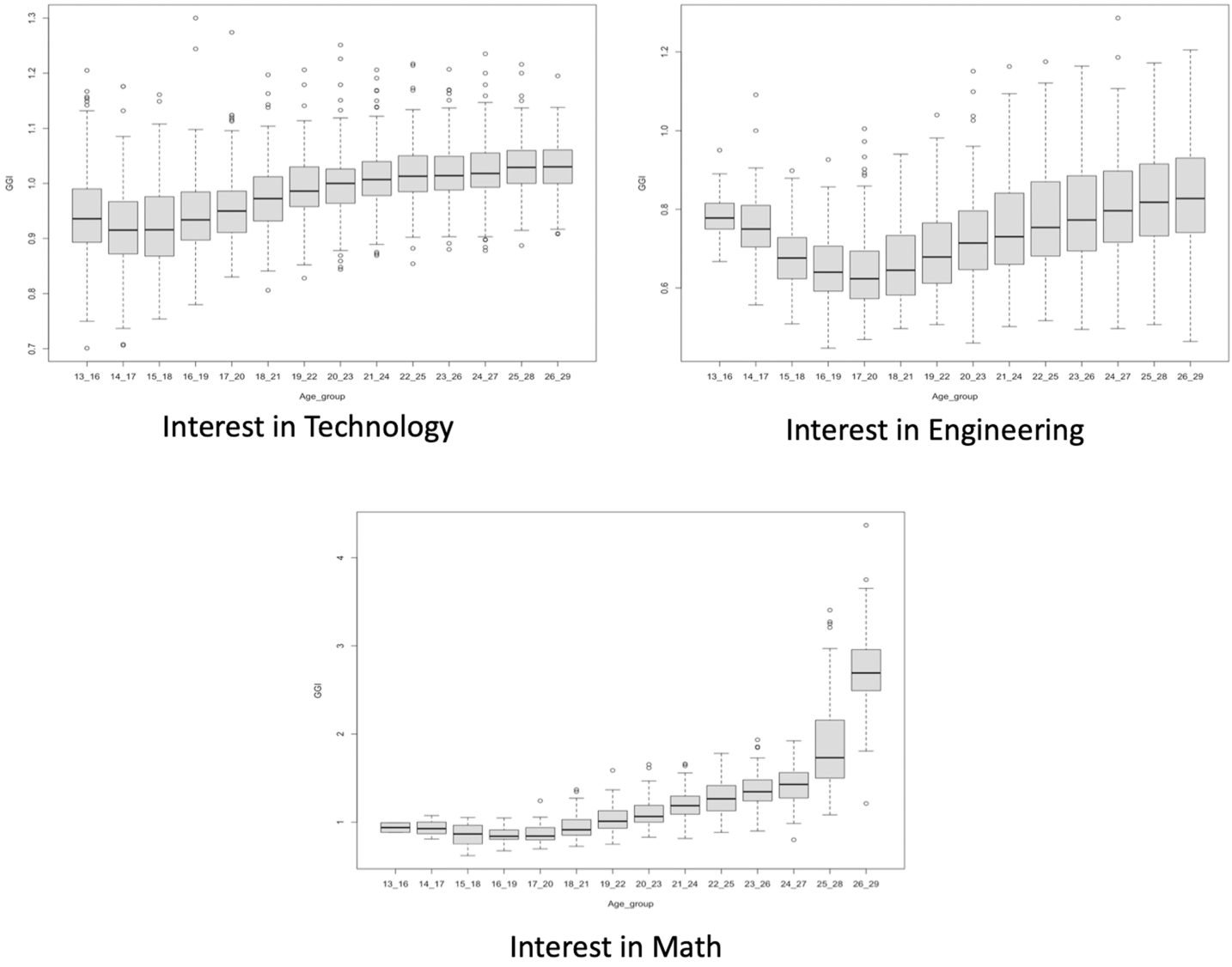


Figure A1 **Facebook GGIs for interests in 'Technology', 'Engineering', and 'Math'.** Full-size DOI: 10.7717/peerj-cs.994/fig-8

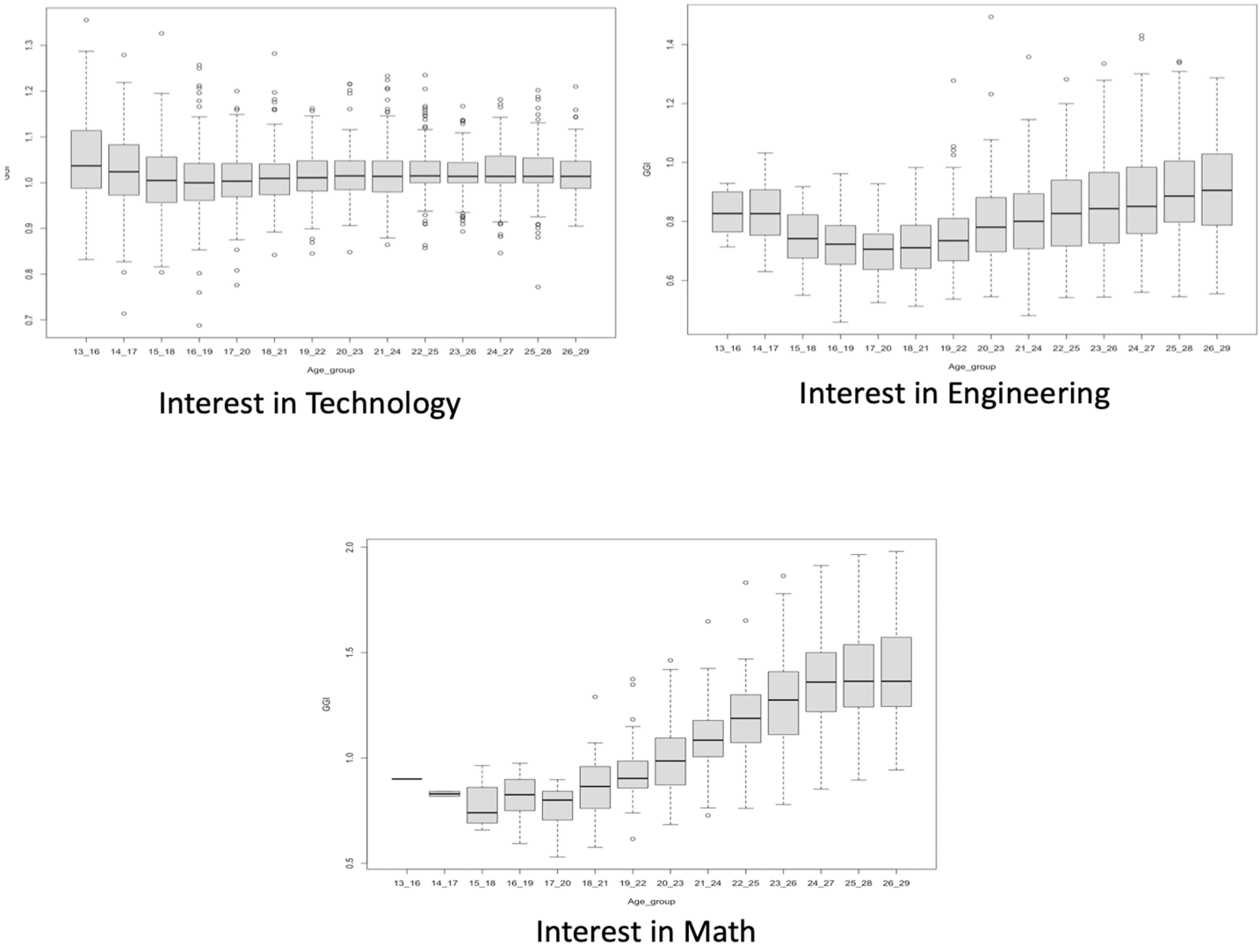


Figure A2 **Instagram GGIs for interests in 'Technology', 'Engineering', and 'Math'.** Full-size DOI: 10.7717/peerj-cs.994/fig-9

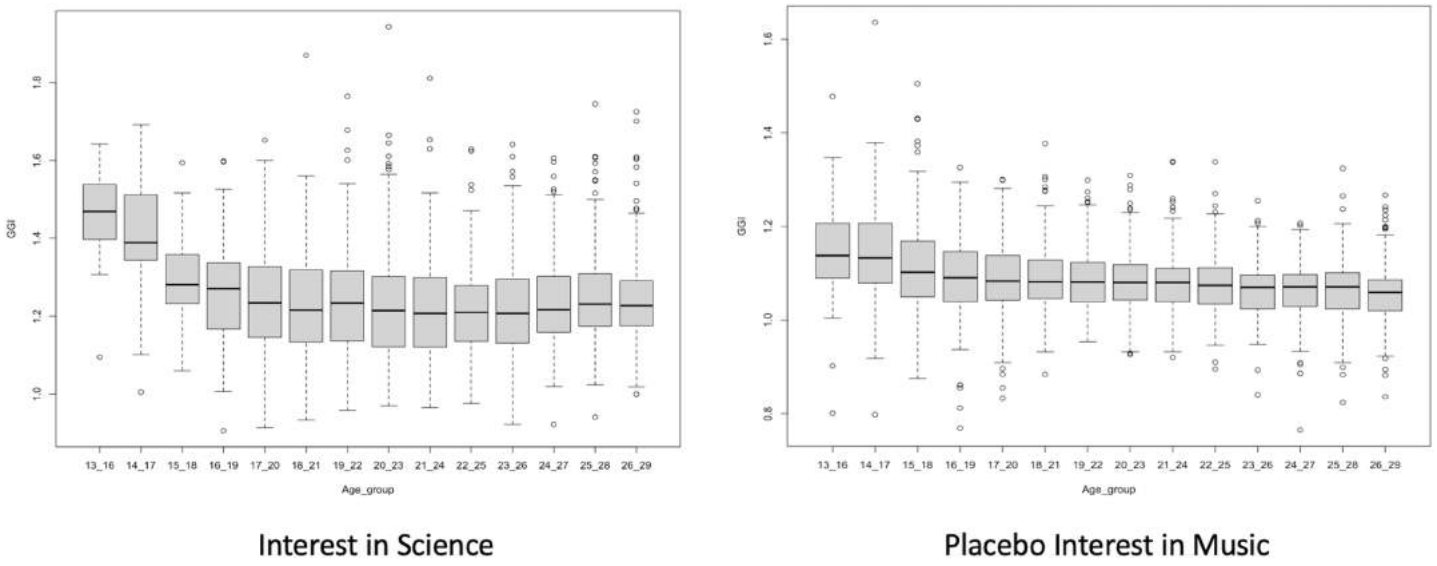


Figure A3 **Side-by-side comparison between the GGIs for interest in 'Science' and GGIs for interest in placebo (music) on Instagram.**
Full-size DOI: 10.7717/peerj-cs.994/fig-10

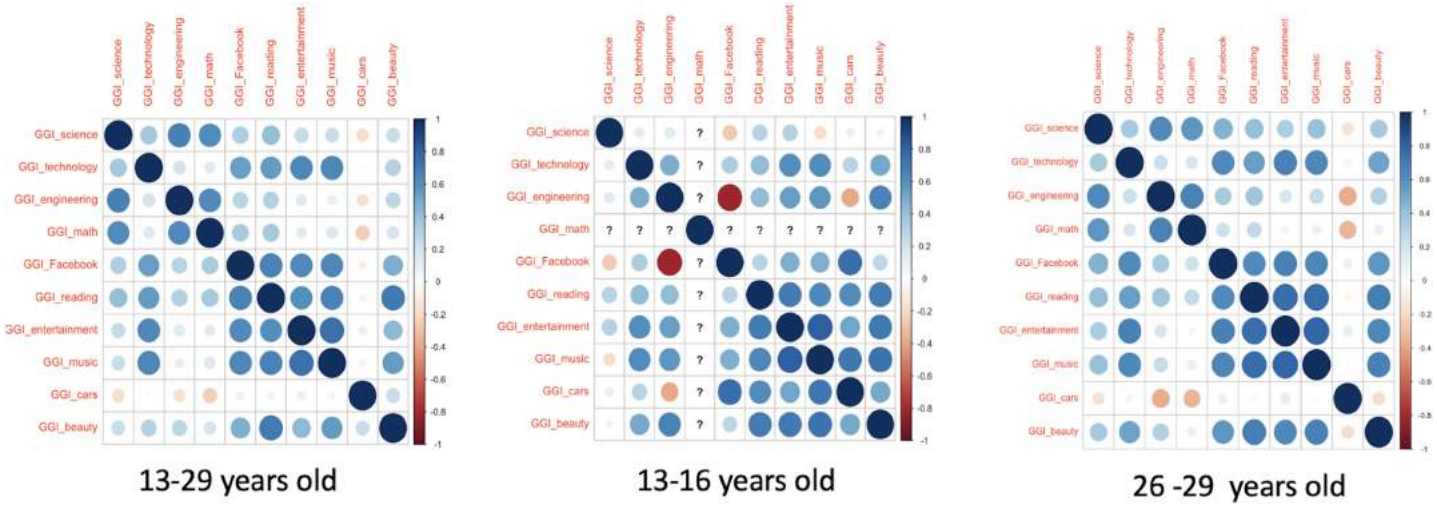


Figure A4 **Correlation between the STEM GGIs and the placebo GGIs for age groups 13–29, 13–16, and 26–29 years old.**
Full-size DOI: 10.7717/peerj-cs.994/fig-11

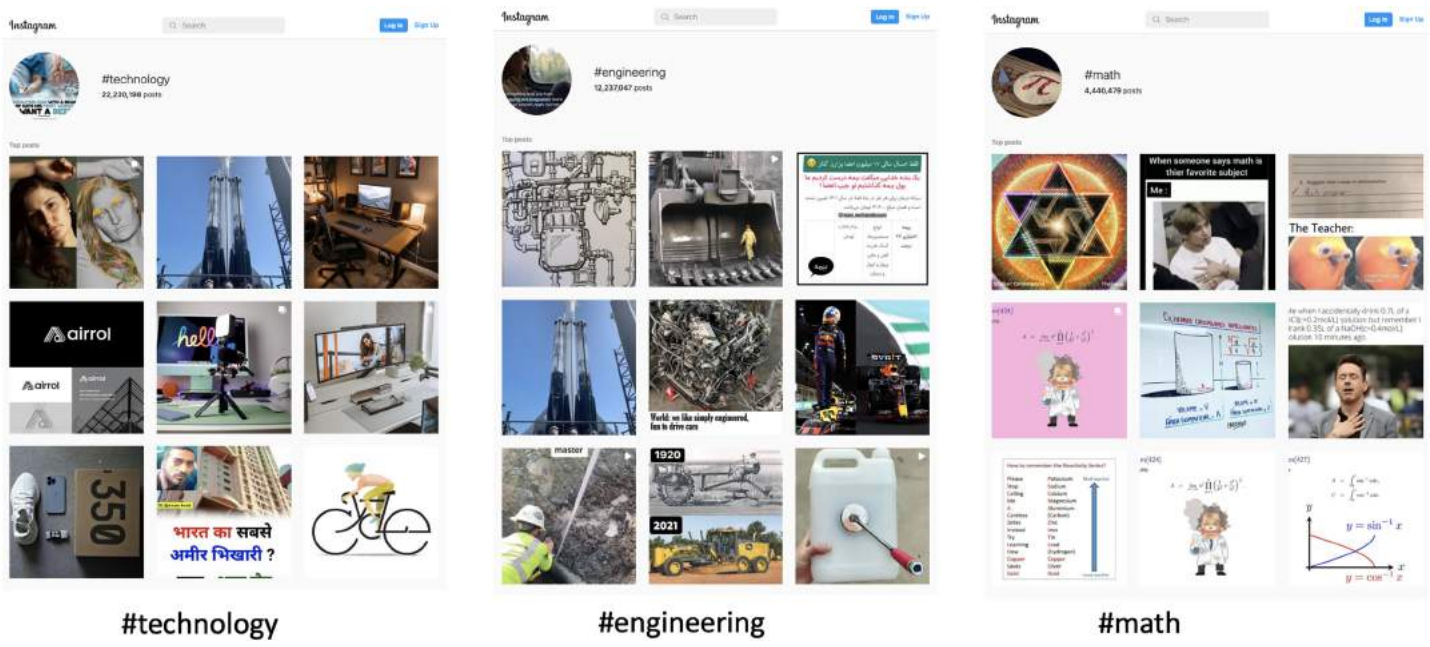


Figure A5 **The hashtags #technology, #engineering, and #math on Instagram.** Full-size DOI: 10.7717/peerj-cs.994/fig-12

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Reham Al Tamime conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Ingmar Weber conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available at figshare: Al Tamime, Reham (2022): PeerJ Article Supplemental Files. figshare. Dataset. https://doi.org/10.6084/m9.figshare.19561288.v1.

REFERENCES

Aguiar L, Peukert C, Schaefer M, Ullrich H. 2022. Facebook shadow profiles. SSRN Scholarly Paper ID 4030019. Rochester, NY: Social Science Research Network. *Available at https://papers.ssrn.com/abstract=4030019* (accessed 03 April 2022).

Araujo M, Mejova Y, Weber I, Benevenuto F. 2017. Using Facebook ads audiences for global lifestyle disease surveillance: promises and limitations. In: *WebSci '17: Proceedings of the 2017 ACM on Web Science Conference*. New York, NY, USA: Association for Computing Machinery, 253–257.

Ayuso N, Fillola E, Masiá B, Murillo AC, Trillo-Lado R, Baldassarri S, Cerezo E, Ruberte L, Mariscal MD, Villarroya-Gaudó M. 2021. Gender gap in STEM: a cross-sectional study of primary school students' self-perception and test anxiety in mathematics. *IEEE Transactions on Education* **64(1)**:40–49 DOI 10.1109/TE.2020.3004075.

Bennaceur A, Cano A, Georgieva L, Kiran M, Salama M, Yadav P. 2018. Issues in gender diversity and equality in the UK. In: *GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 5–9.

Berryman SE. 1983. Who will do science? Trends, and their causes in minority and female representation among holders of advanced degrees in science and mathematics. A special report. *Available at https://eric.ed.gov/?id=ED245052* (accessed 12 February 2022).

Brauner P, Ziefle M, Schroeder U, Leonhardt T, Bergner N, Ziegler B. 2018. Gender influences on school students' mental models of computer science: a quantitative rich picture analysis with sixth graders. In: *GenderIT '18: Proceedings of the 4th Conference on Gender & IT*. New York, NY, USA: Association for Computing Machinery, 113–122.

Card D, Payne AA. 2021. High school choices and the gender gap in STEM. *Economic Inquiry* **59(1)**:8–9 DOI 10.1111/ecin.12934.

Cardador MT, Damian RI, Wiegand JP. 2021. Does more mean less?: interest surplus and the gender gap in STEM careers. *Journal of Career Assessment* **29(1)**:76–97 DOI 10.1177/1069072720930658.

Carrell SE, Page ME, West JE. 2010. Sex and science: how professor gender perpetuates the gender gap. *Quarterly Journal of Economics* **125(3)**:1101–1144 DOI 10.1162/qjec.2010.125.3.1101.

Ceci SJ, Ginther DK, Kahn S, Williams WM. 2014. Women in academic science: a changing landscape. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* **15(3)**:75–141 DOI 10.1177/1529100614541236.

Cheryan S, Ziegler SA, Montoya AK, Jiang L. 2017. Why are some STEM fields more gender balanced than others? *Psychological Bulletin* **143(1)**:1–35 DOI 10.1037/bul0000052.

Cowgill C, Halper L, Rios K, Crane P. 2021. Differential effects of framing the gender gap in STEM recruitment interventions. *Psychology of Women Quarterly* **45(1)**:61–78 DOI 10.1177/0361684320965123.

Delaney JM, Devereux PJ. 2019. Understanding gender differences in STEM: evidence from college applications. *Economics of Education Review* **72(3)**:219–238 DOI 10.1016/j.econedurev.2019.06.002.

Dubois A, Zagheni E, Garimella K, Weber I. 2018. Studying migrant assimilation through Facebook interests. In: Staab S, Koltsova O, Ignatov D, eds. *Social Informatics. SocInfo 2018. Lecture Notes in Computer Science.* Vol. 11186. Cham: Springer DOI 10.1007/978-3-030-01159-8_5.

Faitar G, Faitar S. 2013. Gender gap and stem career choices in 21st century American education. *Procedia-Social and Behavioral Sciences. 4th International Conference on New Horizons in Education* **106**:1265–1270 DOI 10.1016/j.sbspro.2013.12.142.

Garcia D, Kassa YM, Cuevas A, Cebrian M, Moro E, Rahwan I, Cuevas R. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences* **115(27)**:6958–6963 DOI 10.1073/pnas.1717781115.

García-Holgado A, Díaz AC, García-Peñalvo FJ. 2019. Engaging women into STEM in Latin America: W-STEM project. In: *TEEM'19: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: Association for Computing Machinery, 232–239.

García-Holgado A, Mena J, García-Peñalvo FJ, Pascual J, Heikkinen M, Harmoinen S, García-Ramos L, Peñabaena-Niebles R, Amores L. 2020. Gender equality in STEM programs: a proposal to analyse the situation of a university about the gender gap. In: *2020 IEEE Global Engineering Education Conference (EDUCON)*. 1824–1830.

García-Holgado A, Verdugo-Castro S, Sánchez-Gómez MC, García-Peñalvo FJ. 2019. Trends in studies developed in Europe focused on the gender gap in STEM. In: *Interacción'19: Proceedings of the XX International Conference on Human Computer Interaction*. New York, NY, USA: Association for Computing Machinery, 1–8.

Gil-Clavel S, Zagheni E. 2019. Demographic differentials in Facebook usage around the world. In: *Proceedings of the International AAAI Conference on Web and Social Media 13*. 647–650.

Gomez SC, Alvarado LKA, Nisperuza GLB. 2020. Women in STEM: does college boost their performance? *Higher Education: The International Journal of Higher Education Research* **79(5)**:849–866 DOI 10.1007/s10734-019-00441-0.

González-González CS, García-Holgado A. 2021. Strategies to gender mainstreaming in engineering studies: a workshop with teachers. In: *Interacción '21: Proceedings of the XXI International Conference on Human Computer Interaction*. New York, NY, USA: Association for Computing Machinery, 1–5.

Han SW. 2016. National education systems and gender gaps in STEM occupational expectations. *International Journal of Educational Development* **49**:175–187 DOI 10.1016/j.ijedudev.2016.03.004.

Haranko K, Zagheni E, Garimella K, Weber I. 2018. Professional gender gaps across US cities. In: *Proceedings of the International AAAI Conference on Web and Social Media 12.1*.

Heilbronner NN. 2013. The STEM pathway for women: what has changed? *Gifted Child Quarterly* **57(1)**:39–55 DOI 10.1177/0016986212460085.

Hill C, Corbett C, St. Rose A. 2010. *Why so few? Women in science, technology, engineering, and mathematics*. Washington, D.C.: American Association of University Women.

Kashyap R, Verkroost F. 2021. Analysing global professional gender gaps using LinkedIn advertising data. *EPJ Data Science* **10(1)**:39 DOI 10.1140/epjds/s13688-021-00294-7.

Kashyap R, Fatehkia M, Al Tamime R, Weber I. 2020. Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demo-graphic Research* **43(27)**:779–816 DOI 10.4054/DemRes.2020.43.27.

Legewie J, DiPrete TA. 2014. The high school environment and the gender gap in science and engineering. *Sociology of Education* **87(4)**:259–280 DOI 10.1177/0038040714547770.

Levine M, DiScenza DJ. 2018. Sweet, sweet science: addressing the gender gap in STEM disciplines through a one-day high school program in sugar chemistry. *Journal of Chemical Education* **95(8)**:1316–1322 DOI 10.1021/acs.jchemed.7b00900.

Levine M, Serio N, Radaram B, Chaudhuri S, Talbert W. 2015. Addressing the STEM gender gap by designing and implementing an educational outreach chemistry camp for middle school girls. *Journal of Chemical Education* **92(10)**:1639–1644 DOI 10.1021/ed500945g.

Mann A, DiPrete T. 2013. Trends in gender segregation in the choice of science and engineering majors. *Social Science Research* **42(6)**:1519–1541 DOI 10.1016/j.ssresearch.2013.07.002.

Mann A, DiPrete T. 2016. The consequences of the national math and science performance environment for gender differences in STEM aspiration. *Sociological Science* **3**:568–603 DOI 10.15195/v3.a25.

McDaniel A. 2016. The role of cultural contexts in explaining cross-national gender gaps in STEM expectations. *European Sociological Review* **32(1)**:122–133 DOI 10.1093/esr/jcv078.

Mejova Y, Rajiv Gandhi H, Jivanbhai Rafaliya T, Rameshbhai Sitapara M, Kashyap R, Weber I. 2018. Measuring Subnational digital gender inequality in India through gender gaps in Facebook use. In: *COMPASS '18: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. New York, NY, USA: Association for Computing Machinery, 1–5.

Mejova Y, Weber I, Fernandez-Luque L. 2018. Online health monitoring using Facebook advertisement audience estimates in the United States: evaluation study. *JMIR Public Health and Surveillance* **4(1)**:e30 DOI 10.2196/publichealth.7217.

Miller DI, Wai J. 2015. The bachelor's to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis. *Frontiers in Psychology* **6**:37 DOI 10.3389/fpsyg.2015.00037.

Moreno L, González Y, Segura y I, Martínez P. 2014. Women in computer science: survey on the perception of the women's participation in STEM studies. In: *Interacción'14: Proceedings of the XV International Conference on Human Computer Interaction*. New York, NY, USA: Association for Computing Machinery, 1–2.

Obradovich N, Özak Ö, Martín I, Ortuño-Ortín I, Awad E, Ce brián M, Cuevas R, Desmet K, Rahwan I, Cuevas A. 2020. Expanding the measurement of culture with a sample of two billion humans. Working paper 27827. Working Paper Series. National Bureau of Economic Research. *Available at https://www.nber.org/papers/w27827* (accessed 22 December 2021).

Palotti J, Adler N, Morales-Guzman A, Villaveces J, Sekara V, Herranz MG, Al-Asad M, Weber I. 2020. Monitoring of the Venezuelan exodus through Facebook's advertising platform. *PLOS ONE* **15(2)**:e0229175 DOI 10.1371/journal.pone.0229175.

Reinking A, Martin B. 2018. The gender gap in STEM fields: theories, movements, and ideas to engage girls in STEM. *Journal of New Approaches in Educational Research* **7(2)**:148–153 DOI 10.7821/naer.2018.7.271.

Ribaupierre HD, Jones K, Loizides F, Cherdantseva Y. 2018. Towards gender equality in software engineering: the NSA approach. In: *GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 10–13.

Ribeiro FN, Kansaon D, Benevenuto F. 2019. Leveraging the Facebook ads platform for election polling. In: *WebMedia '19: Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 305–312.

Riegle-Crumb C, King B. 2010. Questioning a White male advantage in STEM: examining disparities in college major by gender and race/ethnicity. *Educational Researcher* **39(9)**:656–664 DOI 10.3102/0013189X10391657.

Sabri N, Kashyap R, Weber I. 2021. Examining global mobile diffusion and mobile gender gaps through Facebook's advertising data. In: *HT '21: Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. New York, NY, USA: Association for Computing Machinery, 287–290.

Saha K, Weber I, Birnbaum ML, De Choudhury M. 2017. Characterizing awareness of schizophrenia among Facebook users by leveraging Facebook advertisement estimates. *Journal of Medical Internet Research* **19(5)**:e156 DOI 10.2196/jmir.6815.

Sassler S, Michelmore K, Smith K. 2017. A tale of two majors: explaining the gender gap in STEM employment among computer science and engineering degree holders. *Social Sciences* **6(3)**:69 DOI 10.3390/socsci6030069.

Sax LJ, Newhouse KNS. 2018. Disciplinary field specificity and variation in the STEM gender gap. *New Directions for Institutional Research* **2018(179)**:45–71 DOI 10.1002/ir.20275.

Soylu YN, Adams G. 2020. A cultural psychological model of cross-national variation in gender gaps in STEM participation. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* **24(4)**:345–370 DOI 10.1177/1088868320947005.

Stewart I, Flores RD, Riffe T, Weber I, Zagheni E. 2019. Rock, rap, or reggaeton?: assessing Mexican immigrants' cultural assimilation using Facebook data. In: *WWW '19: The World Wide Web Conference*. New York, NY, USA: Association for Computing Machinery, 3258–3264.

Stoet G, Geary DC. 2018. The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science* **29(4)**:581–593 DOI 10.1177/0956797617741719.

Tandrayen-Ragoobur V, Gokulsing D. 2021. Gender gap in STEM education and career choices: what matters? *Journal of Applied Research in Higher Education* **2(3)**:50 DOI 10.1108/JARHE-09-2019-0235.

Tobar Subía Contento LM, Nohemi Gamez Aparicio B. 2020. The gender gap broad the path for women in STEM. In: *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: Association for Computing Machinery, 187–192.

Torres-Ramos S, Retamoza-Vega PR, Fajardo-Robledo NS, Neri-Cortés C, Rodríguez-Betancourtt VM, Pérez-Carrillo LA. 2020. Towards increasing of STEM-women professionals by implementing projects that reduce the gender gap: a study case in Universidad de Guadalajara. In: *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: Association for Computing Machinery, 166–172.

United Nations. 2022. International day of women and girls in science. United Nations. Publisher: United Nations. *Available at https://www.un.org/en/observances/women-and-girls-in-science-day/* (accessed 12 January 2022).

Verdugo-Castro S, García-Holgado A, Sánchez-Gómez MC. 2019. Analysis of instruments focused on gender gap in STEM education. In: *TEEM'19: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: Association for Computing Machinery, 999–1006.

Verdugo-Castro S, García-Holgado A, Sánchez-Gómez MC. 2020. Interviews of Spanish women in STEM: a multimedia analysis about their experiences. In: *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multi-culturality*. New York, NY, USA: Association for Computing Machinery.

Verdugo-Castro S, Sánchez-Gómez MC, García-Holgado A, Bakieva M. 2020. Pilot study on university students' opinion about STEM studies at higher education. In: *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York, NY, USA: Association for Computing Machinery, 158–165.

Verkroost F, Kashyap R, Garimella VRK, Weber I, Zagheni E. 2020. Tracking global gender gaps in information technology using online data. In: *Digital Skills Insights*. Geneva: ITU Publications, 81–93.

Vieira C, Ribeiro F, Vaz de Melo PO, Benevenuto F, Zagheni E. 2020. Using Facebook data to measure cultural distance between countries: the case of Brazilian cuisine. In: *WWW '20: Proceedings of the Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, 3091–3097.

Vieira C, Vasconcelos M. 2021. Using Facebook ads data to assess gender balance in STEM: evidence from Brazil. In: *WWW '21: Companion Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 145–153.

Wang MT, Degol JL. 2017. Gender gap in science, technology, engineering, and mathematics (STEM): current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review* **29(1)**:119–140 DOI 10.1007/s10648-015-9355-x.

Weeden KA, Gelbgiser D, Morgan SL. 2020. Pipeline dreams: occupational plans and gender differences in STEM major persistence and completion. *Sociology of Education* **93(4)**:297–314 DOI 10.1177/0038040720928484.

Wood J. 2021. How many women work in STEM? World economic forum. *Available at https://www.weforum.org/agenda/2020/02/stem-gender-inequality-researchers-bias/* (accessed 16 December 2021).

Xu YJ. 2008. Gender disparity in STEM disciplines: a study of faculty attrition and turnover intentions. *Research in Higher Education* **49(7)**:607–624 DOI 10.1007/s11162-008-9097-4.

Zhou Y, Fan X, Wei X, Tai RH. 2017. Gender gap among high achievers in math and implications for STEM pipeline. *The Asia-Pacific Education Researcher* **26(5)**:259–269 DOI 10.1007/s40299-017-0346-1.

PeerJ Computer Science

How Twitter data sampling biases U.S. voter behavior characterizations

Kai-Cheng Yang, Pik-Mai Hui and Filippo Menczer

Observatory on Social Media, Indiana University, Bloomington, Indiana, United States

ABSTRACT

Online social media are key platforms for the public to discuss political issues. As a result, researchers have used data from these platforms to analyze public opinions and forecast election results. The literature has shown that due to inauthentic actors such as malicious social bots and trolls, not every message is a genuine expression from a legitimate user. However, the prevalence of inauthentic activities in social data streams is still unclear, making it difficult to gauge biases of analyses based on such data. In this article, we aim to close this gap using Twitter data from the 2018 U.S. midterm elections. We propose an efficient and low-cost method to identify voters on Twitter and systematically compare their behaviors with different random samples of accounts. We find that some accounts flood the public data stream with political content, drowning the voice of the majority of voters. As a result, these hyperactive accounts are over-represented in volume samples. Hyperactive accounts are more likely to exhibit various suspicious behaviors and to share low-credibility information compared to likely voters. Our work provides insights into biased voter characterizations when using social media data to analyze political issues.

Subjects Data Science, Network Science and Online Social Networks
Keywords Twitter, Election, Voter, Bias, Data sampling

INTRODUCTION

In recent years, social media have served as important platforms for news dissemination and public discussion of politics. With more citizens consuming information and actively expressing their political opinions online, political figures, governments, and agencies have started to adopt social media to reach out to the public and amplify their influence (*Jungherr, 2016*). These activities generate huge amounts of valuable data, making it possible to study voter behavior (*Small, 2011*; *Jungherr, 2014*; *Ausserhofer & Maireder, 2013*) and estimate public opinions around political issues (*Diakopoulos & Shamma, 2010*; *Wang et al., 2012*) through computational approaches. Some researchers even attempt to predict election outcomes (*Tumasjan et al., 2010*; *DiGrazia et al., 2013*; *Burnap et al., 2016*), although the reliability of such predictions has been questioned (*Gayo-Avello, Metaxas & Mustafaraj, 2011*).

However, raw social media data might be seriously biased and even lead to false signals. To demonstrate, we use Twitter as an example of social media platforms given its popularity among U.S. politically active citizens. Figure 1 illustrates the composition of different types of accounts that participate in the online discussion about elections. Ideally, we want to focus on the data generated by real-world voters since they determine the outcomes of elections. But the data stream originates from a population that does not

Submitted 15 February 2022
Accepted 10 June 2022
Published 1 July 2022

Corresponding author
Kai-Cheng Yang, yangkc@iu.edu

Academic editor
Diego Amancio

Additional Information and Declarations can be found on page 18

DOI **10.7717/peerj-cs.1025**

Copyright
2022 Yang et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

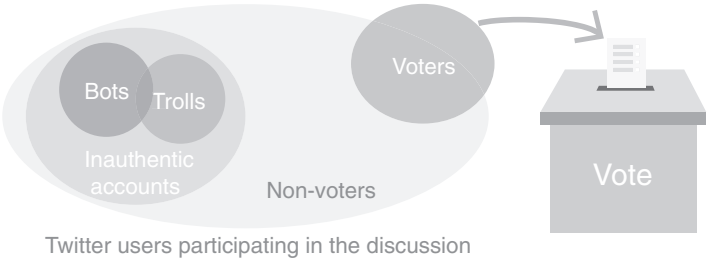


Figure 1 The composition of different sets of accounts participating in election-related discussion on Twitter. The voters determine real-world outcomes, while the online data comes from a population that includes non-voters and inauthentic accounts. The sizes of the sets in the diagram do not correspond to the actual numbers of accounts in each group. Full-size DOI: 10.7717/peerj-cs.1025/fig-1

contain all voters and that contains non-voters and inauthentic accounts. The latter group includes entities like malicious social bots, foreign trolls that have been shown to manipulate public opinion regarding elections, and more (*Bessi & Ferrara, 2016*; *Deb et al., 2019*; *Stella, Ferrara & De Domenico, 2018*; *Ferrara, 2017*; *Badawy, Ferrara & Lerman, 2018*; *Shao et al., 2018*; *Zannettou et al., 2019a*). Relying on content generated by such accounts is clearly problematic.

To understand potentially biased characterizations of U.S. voters based on observations of Twitter data, we propose a method to extract self-identified voters during the 2018 U.S. midterm elections and use the resulting group as a baseline. We also propose two variants of this methods that are more efficient to implement. We further create two different samples of accounts. First, naive sampling based on tweet volume leads to over-representation of hyperactive accounts. Second, random resampling of unique accounts removes this bias. In this exploratory analysis, we systematically compared the accounts selected according to these four different methods to the baseline. We examined account characteristics such as their profiles, activity in the discussion, general sentiment expression, and political alignment. To quantify the prevalence of different questionable behaviors, we also analyzed their sharing of low-credibility information, signs of automation, and so on.

We found that hyperactive accounts are more likely to exhibit suspicious behaviors and share low-credibility information compared to likely voters. These findings further our understanding of biases in characterizations of U.S. voters based on Twitter data, providing useful guidance toward political studies on social media. For journalists and the public at large, it reveals important insights about the accounts they encounter online everyday.

RELATED WORK

Politics-related social media studies

Many recent computational social science and political science studies are dedicated to understanding online behaviors of voters in political discussions. To name a few, *Small (2011)* studies the usage of political hashtags. *Jungherr (2014)* investigates the logic of

political coverage on Twitter by comparing it with traditional media. And *Ausserhofer & Maireder (2013)* study interrelations between the users.

Another line of work seeks to predict election outcomes using indicators estimated from social media data. In the context of a German election, *Tumasjan et al. (2010)* show that the number of tweets can reflect the election result. A study by *DiGrazia et al. (2013)* similarly finds a correlation between the votes and candidate name mentions on Twitter in the 2010 U.S. midterm elections while accounting for other predictors. Considering the same elections, the model proposed by *Livne et al. (2011)* uses variables derived from Twitter content and network features and achieves decent prediction performance. Other studies depend on the sentiment of the tweets. For example, *O'Connor et al. (2010)* found that tweet sentiment is correlated with presidential job approval and consumer confidence polls. *Ceron et al. (2014)* demonstrate the usefulness of sentiment analysis in multiple elections. *Burnap et al. (2016)* show similar results in the 2015 U.K. general election. Despite these efforts, the reliability of electoral predictions based on social media data has been questioned (*Gayo-Avello, Metaxas & Mustafaraj, 2011*).

Questionable actors and behaviors

The politics-related social media studies reviewed above usually assume that social media data comes from real users and barely mention inauthentic accounts. This is not surprising, as studies of inauthentic actors on social media have only started to emerge in recent years. Here we briefly cover the studies about malicious social bots and trolls.

Social bots are social media accounts controlled completely or in part by algorithms (*Stocking & Sumida, 2018*; *Ferrara et al., 2016*). Their automated nature allows a bot creator to easily generate large amounts of content or create false appearance of popularity. Reports show that malicious bots were involved in recent elections in Europe and the U.S. (*Bessi & Ferrara, 2016*; *Deb et al., 2019*; *Stella, Ferrara & De Domenico, 2018*; *Ferrara et al., 2020*; *Gorodnichenko, Pham & Talavera, 2021*). Close analyses suggest that bots tend to have suspicious profile information, recent creation dates, large numbers of tweets, and limited numbers of original posts (*Ferrara et al., 2016*). Questionable behaviors by malicious bots include astroturfing (*Conover et al., 2011*), spreading misinformation (*Shao et al., 2018*), and using inflammatory language to irritate others (*Stella, Ferrara & De Domenico, 2018*).

In this article we use the term "trolls" to refer to fake accounts controlled by state-sponsored agencies. Studies suggest that these accounts have interfered with the 2016 U.S. presidential election (*Badawy, Ferrara & Lerman, 2018*; *Zannettou et al., 2019b*, *2019a*). The characteristics of the trolls can be different from those of random Twitter accounts: they tend to use deceptive language (*Addawood et al., 2019*), their posts come from different Twitter clients, and their creation dates are concentrated (*Zannettou et al., 2019a*).

Note that the bot and troll behaviors mentioned above may only represent a subset of inauthentic behaviors online; other types of suspicious behaviors may remain unnoticed. As a result, our understanding of the prevalence of questionable behaviors among typical Twitter users is limited.

Voter identification

On social media, voters are often identified by conducting surveys. *Mellon & Prosser (2017)* used a nationally representative sample in the U.K. and find that Twitter and Facebook users tend to be younger, better-educated, more liberal, and to pay more attention to politics compared to voters at large. The Pew Research Center publishes a series of studies in which they first survey a group of U.S. adults from a nationally representative panel for their demographics and political ideology. Then they link the survey data to Twitter handles shared by the participants (*Wojcik & Hughes, 2019*; *Hughes et al., 2019*). Among other things, they find that a small portion of users generate most of the political content on Twitter and that politics is not the main topic of conversation for voters.

Another approach starts from social media handles and then attempts to establish links to voters. *Barberá (2016)* proposed a method to match geolocated tweets with voting registration records to obtain the demographics of the people behind the Twitter handles. *Grinberg et al. (2019)* matched Twitter profiles against voter registration information, finding that a small portion of U.S. voters account for the majority of fake news exposure and sharing during the 2016 presidential election.

The approaches introduced above can be powerful in connecting online behaviors and demographics of voters. However, high quality data sources like nationally representative panels, voting registration records, and large-scale surveys are expensive and out of reach for many researchers. Moreover, these approaches can hardly be automated. In their recent work, *Deb et al. (2019)* collected tweets related to the 2018 U.S. midterm elections and treated accounts tweeting the hashtag *#ivoted* on election day as actual voters. This method can be applied to large-scale data efficiently and thus makes rapid or even real-time analysis possible, although we show here that it introduces some bias.

DATA

Tweet collection

In this study, we focus on Twitter data around the 2018 U.S. midterm elections. To create a dataset containing most of the election-related content, we collect tweets that include any hashtag in a query through Twitter's free filtering API. The query includes a list of 143 election-related hashtags from a previous study (*Yang, Hui & Menczer, 2019a*). The list is initialized with several well-known hashtags like *#maga*, *#bluewave*, and *#2018midterms*. A snowball sampling method based on co-occurrence is then applied to expand the list (*Conover et al., 2011*). Hashtags used for each state's Senate election (*e.g.*, *#AZsen* and *#TXsen*) are also added to the list. We then manually check the list and remove irrelevant hashtags. This method is designed to capture most of the election-related tweets with high precision; although the list was not updated during data collection, newly emerging political hashtags likely co-occurred with those already on the list. The collection started on 05 October, 2018 and continued until the end of 2018, resulting in a dataset with about 60 M tweets by 3.5 M unique users.

Naive accounts sampling

Our goal is to understand the characteristics of typical Twitter users and compare them with identified voters. For a meaningful comparison, we used the collected dataset as the sampling pool, *i.e.*, we focus on accounts who discussed the 2018 U.S. midterm elections. We deployed two different methods to sample typical accounts. With an unweighted sampling method, we simply select unique accounts in our dataset randomly, hoping to get a representative sample. We call the group of accounts obtained with this sampling method `ByAccount`.

Sampling users from the tweets in our dataset favors active accounts: the chance of being selected is proportional to the number of tweets each account has in our dataset. We call the group of accounts sampled by this method `ByTweet`. In the following analyses, we will show that `ByTweet` accounts are hyperactive and generate disproportionately more content. As a result, Twitter users are more likely to be influenced by these accounts directly or indirectly, and content analysis based on the Twitter stream could be biased accordingly. The `ByAccount` and `ByTweet` groups have the same size as the `Voter` group, described next.

IDENTIFICATION OF VOTERS

In this section we present high-precision methods to identify likely voters on Twitter using the data collected. These accounts are used as a reference group to describe authentic behaviors in the online discussion regarding U.S. politics.

The basic idea is to use the messages and pictures posted by users on election day. We followed the approach of *Deb et al. (2019)* and used the hashtag *#ivoted* as a first filter. The hashtag was promoted by various social media platforms to encourage people to vote in the midterm elections, and it became quite popular for users to tweet about it after voting on election day (06 November, 2018). However, considering all accounts tweeting the *#ivoted* hashtag as voters may result in false positives since inauthentic accounts can also post the hashtag. In fact, *Deb et al. (2019)* exclude about 20% of the accounts that tweeted *#ivoted* in their analysis, based on high Botometer scores (*Yang et al., 2019b*).

Instead of attempting to exclude all problematic accounts from those tweeting *#ivoted*, which is itself prone to false positive and false negative errors, we aimed to include only those who are very likely to be real voters. This approach yields voter accounts with high precision at the expense of recall, but this is acceptable for our purpose of characterizing likely voters. We notice that many voters shared pictures of themselves wearing "I Voted" stickers or buttons after voting (see Fig. 2). Based on this observation, we apply a second filter that verifies whether the *#ivoted* hashtag is accompanied by a picture showing a real person having voted.

Manual annotation

To identify the voters in our dataset, we first extract all the tweets posted on election day that include both the *#ivoted* hashtag and images. We download all the pictures for manual annotation. A single criterion is used in the annotation: whether we can reach the conclusion that a real person just voted based on the picture. The annotations based on this



Figure 2 Examples of images providing positive (top) and negative (bottom) cases in our manual annotation of voters. Images are cropped for privacy consideration.
Full-size DOI: 10.7717/peerj-cs.1025/fig-2

criterion turn out to be quite straightforward. In most of the positive cases, the picture clearly shows a person with an "I Voted" sticker or button on their clothes or body. There are also some borderline cases where the pictures show voting-related memes or computer-generated "I Voted" stickers. Although the corresponding users might have voted, we still consider these as negative cases because sharing these pictures can be faked easily. Figure 2 shows examples of both positive and negative cases.

Although *#ivoted* is not on the list of hashtags used for collecting the tweets, it often co-occurs with other hashtags in the query, so we end up with 40,078 tweets containing this hashtag. Among these, 13,084 have pictures attached. Through annotation and data cleaning, we identify 8,092 voters. We call this group of accounts `Voter`. For comparison, we also sample an equal number of accounts that tweeted *#ivoted* with no pictures and call this group `Ivoted`.

The free Twitter API is sufficient for collecting the data. Manual annotation is relatively time-consuming: it took about 30 h for the first author to go through all the images. But the task is relatively straightforward, so no special training is needed. Moreover, implementing this method only requires basic programming skills and a Twitter developer account.

Automated annotation

Since the human annotation is the bottleneck of the method described above, we also implemented an automated annotation process. The key challenge is to translate the annotation criterion into a format that machines can understand. Since most of the positive cases have the text "vote" in the picture, we operationalize the annotation criterion

Table 1 Summary of the account groups examined in this study.

Account group	Size	Methodology
`Voter`	8,092	Manually-annotated likely voters from accounts tweeting *#ivoted* and photos
`Voter_auto`	5,920	Machine-identified likely voters from accounts tweeting *#ivoted* and photos
`Ivoted`	8,092	Accounts sampled randomly among those tweeting *#ivoted* but no photos
`ByTweet`	8,092	Accounts sampled based on tweet volume in the dataset
`ByAccount`	8,092	Accounts sampled randomly among all those in the dataset

as identifying whether the picture contains the string "vot" (to match "vote," "voted," and "voting").

Traditional optical character recognition software (https://opensource.google/projects/tesseract) does not provide acceptable performance. Therefore we resort to an off-the-shelf deep learning model provided by the Google Cloud service (https://cloud.google.com/vision/docs/ocr). We send all images to the the text detection API endpoint to extract text strings and check whether "vot" is in the response.

In our experiment, the automated approach annotates 46.5% of the pictures as positive. By treating the manual annotations as ground truth, we measure the performance of the automated approach: it achieves 0.81 in precision and 0.59 in recall. Close inspection of misclassified data reveals that most false positive cases are just like the ones in the bottom row of Fig. 2, as expected. Typical false negative cases, on the other hand, are mostly due to limitations of the API in handling blurred, distorted, and rotated pictures. For example, the API fails to detect the text in the upper right image in Fig. 2. We collect the accounts labeled as voters by the automated method and call the group `Voter_auto`.

With the help of Google Cloud or similar services, the whole voter identification process can be automated. The method is therefore fast and cheap; in our experiment, it takes less than 15 min and about $20 to annotate all the pictures. In summary, it is possible to setup an infrastructure to automatically identify likely voters on Twitter in real time at a very low price, if low recall is acceptable. With the rapid development of deep learning technologies, we expect to see an increase in performance and a decrease in price over time.

Table 1 summarizes the size of and methodology used to generate each group: `Voter`, `Voter_auto`, `Ivoted`, `ByTweet`, and `ByAccount`. As mentioned earlier, `Voter` contains 8,092 likely voters and we sample the same amount of accounts in `Ivoted`, `ByTweet`, and `ByAccount` while ensuring that they are mutually exclusive. `Voter_auto` and `Voter` share 4,852 accounts since they are generated from the same pool of accounts tweeting both *#ivoted* and images.

We share the tweet IDs of the collected tweets and the user IDs of the users in different samples in a public data repository (https://doi.org/10.5281/zenodo.5866751). Sharing the raw tweets is prohibited by Twitter's policy. But readers can use the IDs in our dataset to re-hydrate the data and reproduce the analysis.

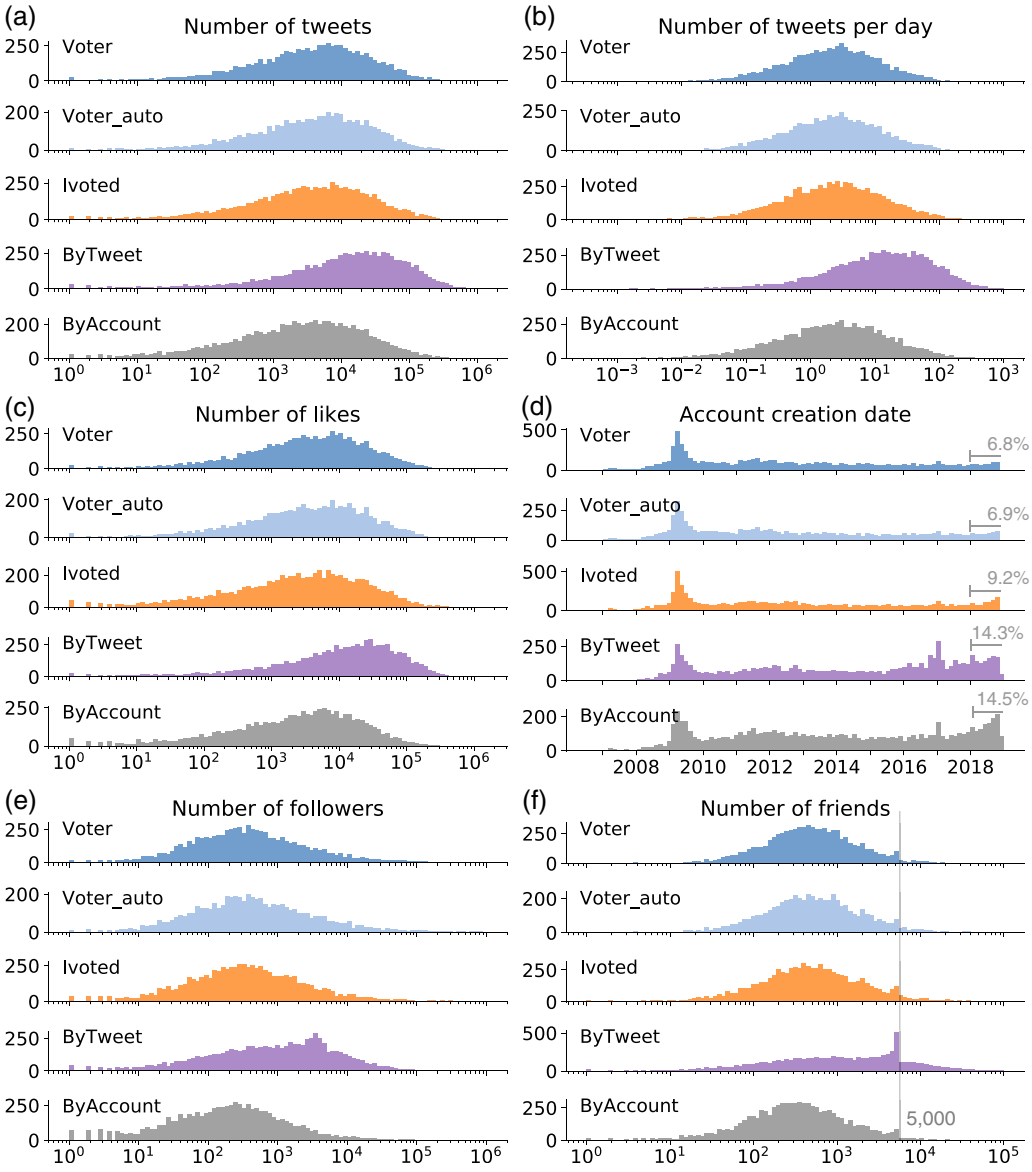


Figure 3 Distributions of (A) number of tweets; (B) number of tweets per day; (C) number of likes; (D) account creation date; (E) number of followers; and (F) number of friends for different groups. The annotations in (D) represent the percentage of accounts that were created in 2018 for each group. The line in (F) indicates 5,000 friends. Full-size DOI: 10.7717/peerj-cs.1025/fig-3

CHARACTERIZATION OF ACCOUNT BEHAVIORS

In this section, we compare various characteristics of accounts in each group listed in Table 1. Our analysis covers account profiles, tweeting behaviors during the elections, sentiment expressions, and political alignment.

Account profiles

We systematically compare the profiles of the accounts in different groups and show selected results in Fig. 3. The information can be accessed directly from the user object embedded in each tweet, which reflects the status of the account at the time the tweet was

collected. Since these user objects may change over time, we use the earliest version of each account in the collection for our comparison.

We show the distributions of total number of tweets in Fig. 3A. It appears that `ByTweet` accounts generate more tweets than other groups across their life time. Since the total number of tweets depends on the account age, we also show the distributions of number of tweets per day in Fig. 3B. This is calculated by dividing an account's total number of tweets by its age, which is defined as the time difference (number of days) between the date of the data collection and the account creation date. Again, we can see that `ByTweet` accounts tweet faster than other groups. Over a half of the `ByTweet` accounts post more than 10 tweets per day; some post more than 100 tweets per day. The distributions of the two measures for other account groups are similar.

We also examine the number of likes and plot the distributions in Fig. 3C. Again, `ByTweet` accounts have more likes than other groups, which have identical distributions. The distributions of the number of likes per day (not shown) suggest that `ByTweet` accounts produce likes faster as well.

The distributions of account creation time are presented in Fig. 3D. Across all groups, we observe a peak around 2009, which is likely caused by the fast growth of Twitter during that year. `ByTweet` and `ByAccount` groups have more accounts created in 2018—right before the midterm elections—than other years. We annotate the percentage for each group in the figure to confirm this observation.

Figures 3E and 3F show the distributions of the numbers of followers and friends, respectively. The number of friends indicates how many other accounts each account is following. We find that `ByTweet` accounts tend to have more followers and friends than other groups. Interestingly, `ByTweet` has more accounts with about 5,000 friends. This is due to an anti-abuse limitation of Twitter, which states that an account cannot follow more than 5,000 friends unless it has more than 5,000 followers (https://help.twitter.com/en/using-twitter/twitter-follow-limit). The pattern here suggests that accounts in `ByTweet` are eagerly expanding their friend lists, until hitting the limit. `ByAccount` accounts tend to have fewer followers and friends than the voter groups.

Through the analyses of the account profiles, we find that `ByTweet` accounts are not only hyperactive in terms of tweeting, as expected given the sampling method; they are also hyperactive in liking and building social networks, and they tend to be more recent.

Tweeting behaviors during elections

We examined the level of participation in the online discussion about the midterm elections for different groups. This is measured by the number of tweets each account has in our dataset (see the distributions in Fig. 4A). The three voter groups have distributions similar to `ByAccount`, with accounts typically having fewer than a hundred political tweets. Due to the sampling method, it is expected that `ByTweet` accounts are more active in the discussion. The discrepancy in political activity is huge: over 65% of the `ByTweet` accounts have more than a hundred political tweets, compared to 12% or less in other groups.

These distributions suggest that the majority of the political tweets are generated by a small group of hyperactive users, which is in line with previous findings (*Hughes et al.,*

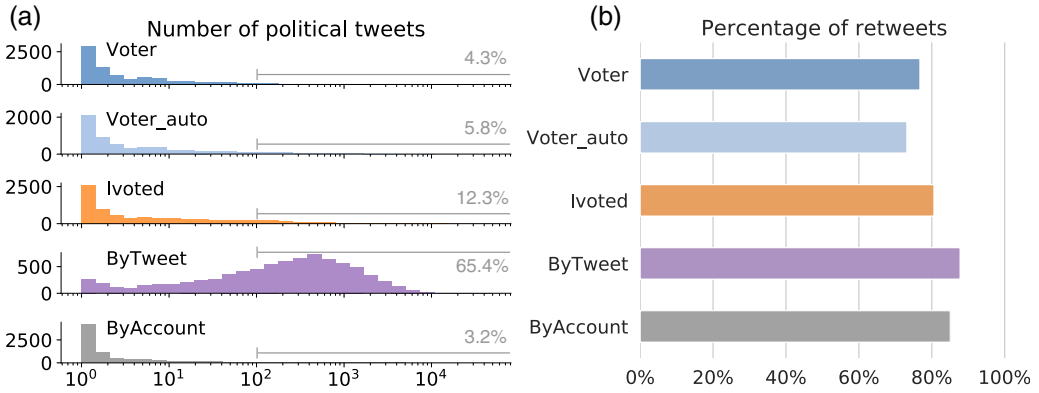


Figure 4 Tweeting behaviors during elections. (A) Distributions of the number of political tweets for different groups. Annotations show percentages of accounts with over 100 political tweets. (B) Percentage of retweets among the tweets generated by different groups. Full-size DOI: 10.7717/peerj-cs.1025/fig-4

2019). This is concerning when one considers that the volume of tweets is used to determine trending topics and viral content. Hyperactive accounts can flood the public data stream with content, drowning the voice of the majority of users, and manipulating engagement metrics and mechanisms that rely on these metrics, such as feed ranking and trending algorithms.

In addition to the absolute number of political tweets, we also examined the retweeting behaviors of different groups. On Twitter, one can amplify messages from other accounts by retweeting or quoting them. Quoting is relatively rare compared to retweeting, so we merge the two to simplify our analysis. For the political tweets generated by each group, we calculate the percentage of retweets and show the results in Fig. 4B. The majority of the tweets in each group are retweets, with `ByTweet` having a particularly high percentage.

Sentiment analysis

Sentiment analysis is commonly adopted in mining social media for public opinions on political matters (*Wang et al., 2012*; *Ramteke et al., 2016*; *Burnap et al., 2016*). Here we aim to compare the sentiment expressions of different groups. We use VADER, an efficient lexicon- and rule-based sentiment analysis tool that works well on Twitter data (*Hutto & Gilbert, 2014*). In VADER's lexicon dictionary, each word and symbol is assigned a sentiment score through crowdsourcing. When applied to a tweet, VADER estimates the negative, positive, and neutral expressions to generate an overall sentiment score.

We show the tweet-level sentiment score distributions for different groups in Fig. 5. A score of zero is assigned to a tweet if it has no sentiment-expressing words. The percentage of neutral tweets is consistent across the groups (about 30%), so we excluded them from Fig. 5 to emphasize the patterns of positive and negative expressions. We find that the general sentiment is consistent across groups.

Political alignment

Since we are interested in behaviors regarding the U.S. elections, it is important to estimate the political alignment of accounts. Following previous studies, we rely on the content

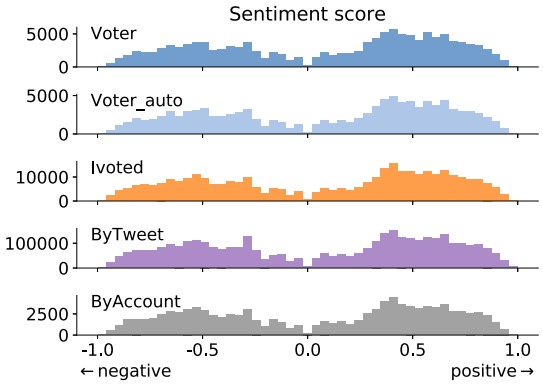


Figure 5 Tweet-level sentiment score distributions for different groups. Positive scores mean positive sentiment and *vice versa*. Sentiment-neutral tweets are excluded from this figure.
Full-size DOI: 10.7717/peerj-cs.1025/fig-5

produced by each account for this analysis. We apply two independent methods, one using hashtags and one using URLs (the links in the tweets) (*Chen et al., 2021*). In both methods, a political alignment score is assigned to each entity (hashtag or URL) in the tweets. The alignment score of each tweet is then obtained by averaging the associated entity scores, and the average score across an account's tweets is used as the alignment score of that account. Next, we briefly introduce how each method assigns political alignment scores to the entities.

Hashtags are widely adopted by Twitter users to label topics and political alignment for their tweets. Therefore it is common to use hashtags to infer user alignment (*Conover et al., 2011*; *Cota et al., 2019*). Instead of manually labeling hashtags, we apply a recently developed approach to obtain continuous alignment scores in a semi-supervised fashion (*Chen et al., 2021*). We first deployed the *word2vec* algorithm (*Mikolov et al., 2013a*, *2013b*) to infer vectorized representations of the hashtags based on their co-occurrence. Since the vectors encode the semantic relations between the entities, the axis between a pair of carefully selected hashtags can represent the political spectrum (*An, Kwak & Ahn, 2018*; *Kozlowski, Taddy & Evans, 2019*). The projected position of a hashtag on this axis reflects its alignment. In this study, we used `#voteblue` and `#votered` as the two anchors of the political alignment axis. We then calculated the relative positions of the remaining hashtags on the axis to produce alignment scores in [−1,1] (negative values mean left-leaning).

Twitter users also commonly use links to share news and information from other platforms. Many websites, especially news outlets, have known political alignment, making it possible to infer an account's political score based on shared links. We adopt a list of 500 news sources with political alignment scores in [−1,1] (negative values mean left-leaning) from previous work (*Bakshy, Messing & Adamic, 2015*). When processing the tweets, we extract the URLs, expand the shortened ones, obtain the domains, and then assign them scores from the list. Entities and tweets with no score assigned are excluded in the evaluation.

We show the distributions of political alignment scores estimated using the two methods in Fig. 6. The two methods generate quantitatively different results but

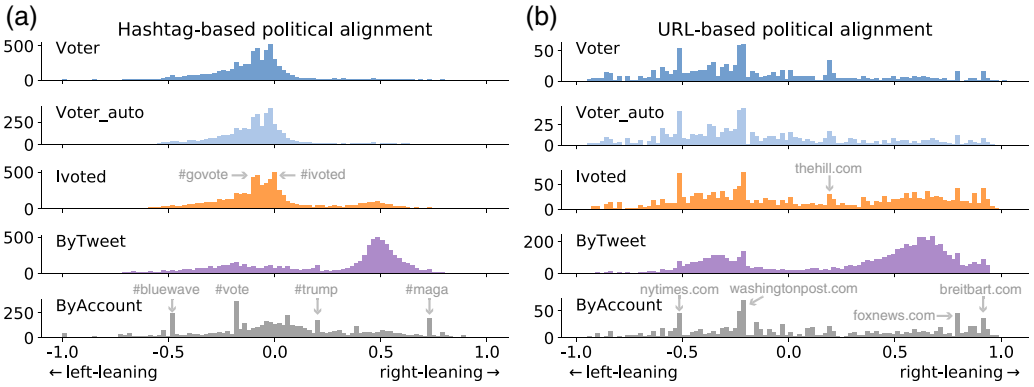


Figure 6 Distributions of political alignment scores for different groups using the (A) hashtag-based and (B) URL-based method. Peaks are due to popular entities, some of which are annotated.
Full-size DOI: 10.7717/peerj-cs.1025/fig-6

qualitatively similar patterns. They suggest that voter accounts tend to be left-leaning, in agreement with previous surveys (*Mellon & Prosser, 2017*; *Wojcik & Hughes, 2019*). `Ivoted` exhibits a similar distribution with `Voter` except that it has more right-leaning accounts. The `ByAccount` group presents a relative diverse and symmetrical distribution across the spectrum. Most surprisingly, accounts in the `ByTweet` group show a bimodal distribution with more accounts on the conservative side. This is inconsistent with the other groups, suggesting that the activity-weighted sampling bias completely distorts the political representation of Twitter accounts.

CHARACTERIZATION OF QUESTIONABLE BEHAVIORS

In this section, we focus on the questionable behaviors of the accounts in different sample groups. Specifically, we examine their automated activities, abnormal tweet deletions, sharing of low-credibility information, and account suspension statuses.

Automated activities

As discussed in the introduction, various inauthentic accounts might participate in online discussions about elections. Therefore it is important to estimate the prevalence of these accounts in different groups. Since social bots are the only type whose prevalence can be statistically estimated using off-the-shelf detection tools, we focus on automated accounts in this subsection.

We adopted the bot detection model proposed by *Yang et al. (2020)*. By strategically selecting a subset of the training dataset, the model achieves high accuracy in cross-validation as well as cross-domain tests. The classifier produces a bot score between zero and one for each account, where larger scores mean more bot-like. The bot score distributions for different groups are shown in Fig. 7. We can see that the `Voter` group has the lowest bot scores, while accounts in `ByTweet` tend to be more bot-like.

Since the majority of political tweets in our dataset are retweets, we also checked whether the sampled accounts amplify messages from likely bots. We called the accounts being retweeted "tweeters" and the accounts amplifying the messages "retweeters." For

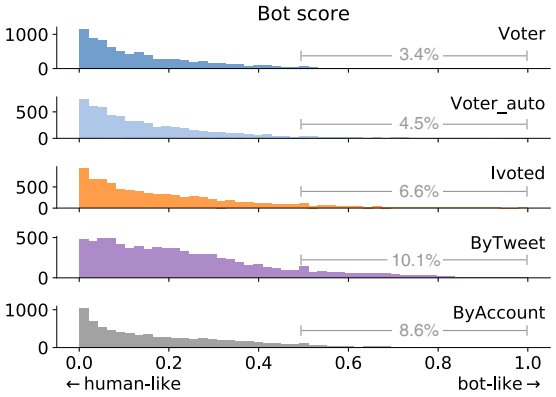


Figure 7 Bot score distributions of different account groups. We annotate the percentage of accounts having bot score above 0.5. Full-size DOI: 10.7717/peerj-cs.1025/fig-7

each retweeter account in the samples, we calculated the average tweeter bot scores weighted by the number of retweets. We plotted the joint distributions of average bot scores of tweeters and bot scores of retweeters in different groups in Fig. 8.

We see that the accounts amplified by the `ByTweet` group tend have higher average bot scores, indicating more bot-like behaviors. On the other hand, `Voter`, `Voter_auto`, and `ByAccount` groups seem to mainly engage human-like accounts. `Ivoted` lies in the middle. We also observe increased density in the top-right region (highlighted by an oval) in Fig. 8C, suggesting interactions among bot-like accounts in the `ByTweet` group.

Abnormal tweet deletion

Twitter users have the right to remove their content and may do so for various legitimate reasons, from fixing typos to removing regrettable content (*Almuhimedi et al., 2013*; *Zhou, Wang & Chen, 2016*). However, this feature can be abused. For example, some politicians delete old tweets for public image management (*Meeks, 2018*). Trolls and bots may systematically delete their tweets to make their actions and intentions harder to detect (*Zannettou et al., 2019a*, *Yang et al., 2019b*).

Since we did not explicitly collect deletion events, we inferred them from our data. User objects that come with the tweets are snapshots of the account profiles at the time of tweet collection. A decrease in the total number of status counts between two consecutive tweets in the collection indicates a tweet deletion event. Note that not all deletion events can be detected with this method. We recorded the maximum drop in status counts for each account.

To analyze abnormal tweet deletion behaviors, we first need to establish a reference deletion frequency. We adopted numbers from a previous study showing that Twitter users delete seven to 11 tweets on average in a week (*Almuhimedi et al., 2013*). In our case, the data collection continued for about three months, so we consider a status count drop of 100 or more tweets as a conservative threshold of abnormal behavior. Figure 9 shows the percentage of accounts with abnormal tweet deletion behaviors in each group. The `ByTweet` group has a much higher ratio of accounts that exhibit abnormal tweet deletion behaviors; this result is robust with respect to higher thresholds.

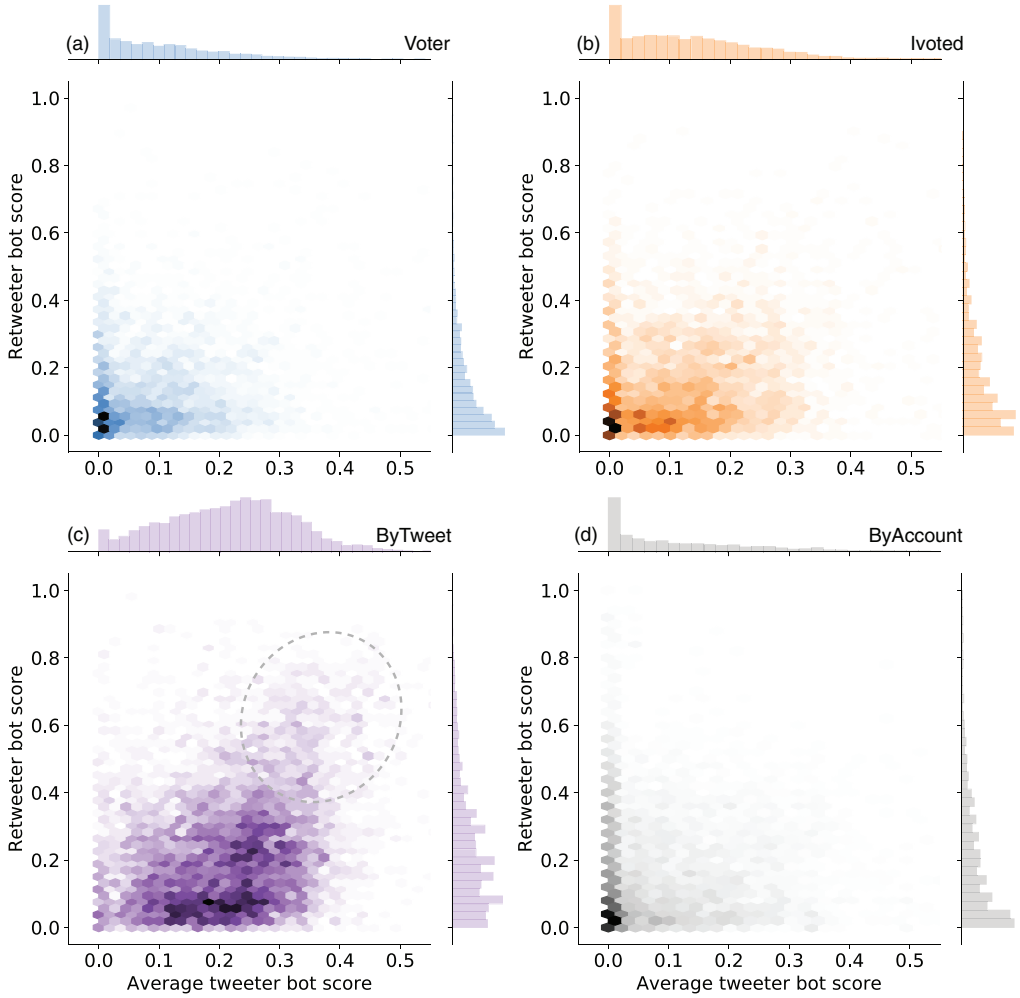


Figure 8 Joint distributions of the average tweeter bot scores and retweeter bot scores in (A) `Voter`, (B) `Ivoted`, (C) `ByTweet`, and (D) `ByAccount` groups. The plot for the `Voter_auto` group is identical to the `Voter` plot and omitted. Accounts without any retweets are ignored in this analysis.
Full-size DOI: 10.7717/peerj-cs.1025/fig-8

Accounts suspension

In the 2018 midterm election season, Twitter reinforced their actions against suspicious activities and suspended many accounts that violated their policy (https://blog.twitter.com/en_us/topics/company/2019/18_midterm_review.html). In line with previous studies (*Alothali et al., 2018*), we consider being suspended by the platform as a signal of illicit behaviors. We checked the status of the accounts on 05 January, 2019 through the Twitter user lookup API and plot the percentages of inaccessible accounts in different groups in Fig. 10. Being inaccessible means the target account is either suspended, deleted, or protected. Users can delete or protect their accounts, so inaccessible accounts are not necessarily suspicious. However, such actions are very rare compared to suspension, so we use inaccessibility as a proxy for suspension.

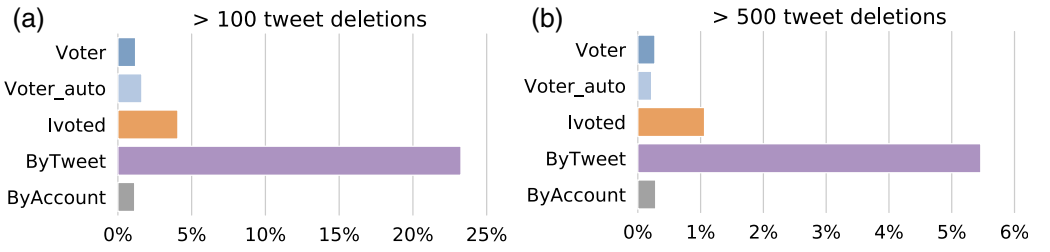


Figure 9 **Percentages of accounts that delete more than (A) 100 and (B) 500 tweets at once in different groups.** Full-size DOI: 10.7717/peerj-cs.1025/fig-9

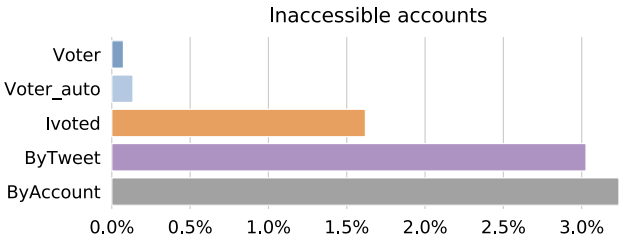


Figure 10 **Percentages of accounts that were inaccessible as of 05 January, 2019 in different groups.** Full-size DOI: 10.7717/peerj-cs.1025/fig-10

Interestingly, `ByAccount` has the highest suspension rate this time followed closely by `ByTweet`. This suggests that from Twitter's policy standpoint, hyperactivity is neither a flag of illicit behavior nor correlated with illicit behaviors. Accounts in `Ivoted` also suffer from suspensions to some extent, while most accounts in `Voter` and `Voter_auto` are not affected.

Sharing low-credibility information

Concern about low-credibility information spreading on social media has been growing ever since the 2016 U.S. presidential election. It is therefore interesting to evaluate the level of involvement for accounts in different groups. We focused on the URLs extracted from the collected tweets and evaluated the credibility of the content at the domain level following the literature (*Shao et al., 2018*; *Grinberg et al., 2019*; *Guess, Nagler & Tucker, 2019*; *Pennycook & Rand, 2019*; *Bovet & Makse, 2019*; *Vosoughi, Roy & Aral, 2018*). We compiled a list of low-credibility domains from recent research papers. A domain was labeled as low-credibility if it is either (a) labeled as low-credibility by *Shao et al. (2018)*; (b) labeled as either "fakenews" or "hyperpartisan" by *Pennycook & Rand (2019)*; (c) labeled as "fakenews," "extremeright," or "extremeleft" by *Bovet & Makse (2019)*; or (d) labeled as "Black," "Red," or "Satire" by *Grinberg et al. (2019)*. This procedure yielded 570 low-credibility domains. Note that we treat hyperpartisan domains as low-credibility.

We followed the same procedure as in the political alignment estimation to obtain the URL domains from each tweet. We excluded from the analysis URLs linking to Twitter itself and other social media like Facebook and YouTube. Most other domains were news outlets. With the help of the list, we are able to characterize the sharing of low-credibility information by the accounts.

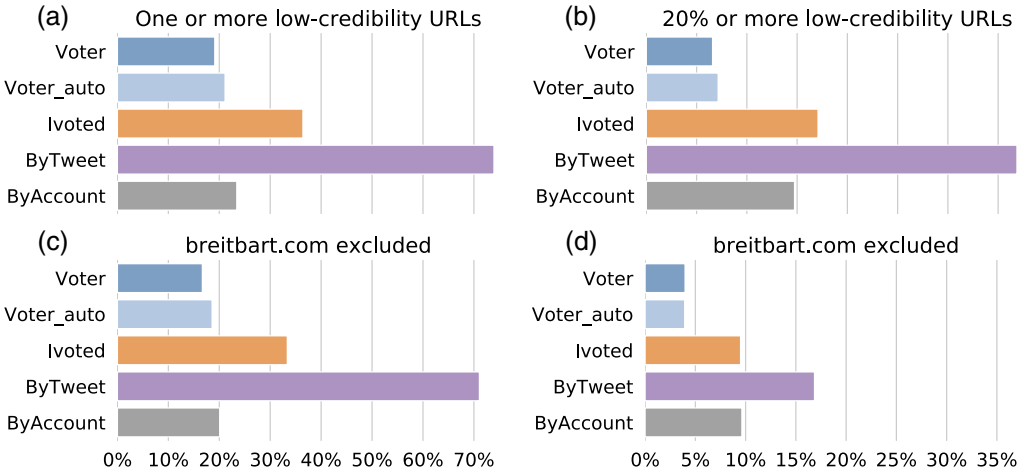


Figure 11 Sharing links to low-credibility sources. (A) Percentages of accounts in different groups sharing at least one URL from low-credibility sources. (B) Percentages of accounts in different groups having at least 20% of shared URLs from low-credibility sources. (C) Same as (A) but breitbart.com is excluded. (D) Same as (B) but breitbart.com is excluded.

Full-size DOI: 10.7717/peerj-cs.1025/fig-11

First, we calculated the percentage of accounts that shared at least one low-credibility URL for each group and plot the results in Fig. 11A. We found that the number for `ByTweet` is alarmingly high—over 70% of the accounts had shared links to low-credibility sources. The other groups have lower percentages. We noticed that breitbart.com appears much more frequently than other low-credibility domains in our dataset. To make sure the pattern observed in Fig. 11A is not dominated by one hyperpartisan source, we performed a robustness test by excluding this domain and show the results in Fig. 11C. The proportions decrease slightly for all groups, but the results were unaffected.

The higher portion of links to low-credibility sources shared by the `ByTweet` group could be trivially due to the higher number of tweets produced by accounts in this group. To exclude this possibility, let us also consider the proportion of low-credibility URLs shared by each account. In Fig. 11B, we show the percentages of accounts with more than 20% of their shared links from low-credibility sources. The results were qualitatively similar, but the percentage of accounts reduces by about half. We also performed the same robustness check by excluding breitbart.com and plot the results in Fig. 11D. The percentages drop further, but the general pattern is similar.

In addition to treating each group as a whole, we can break them down by political alignment. We reported the results for left- and right-leaning accounts from each group in Fig. 12. By comparing accounts with the same political alignment across the groups, we find that the same patterns as before: `ByTweet` accounts have the highest rate of low-credibility content. Within each group, the right-leaning accounts were more likely to share links to low-credibility sources. We performed the same analysis using the proportion of low-credibility URLs (not shown), with identical observations.

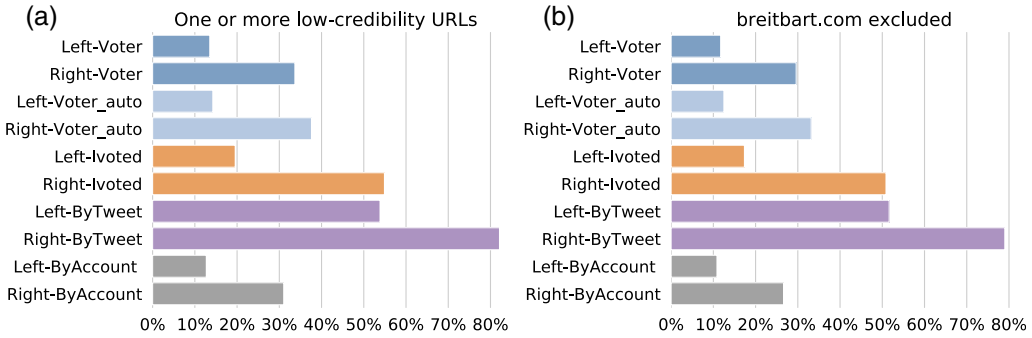


Figure 12 Sharing links to low-credibility sources, by political alignment. (A) Percentages of accounts in different sub-groups that shared at least one URL from low-credibility sources. (B) Same as (A) but breitbart.com is excluded. Full-size DOI: 10.7717/peerj-cs.1025/fig-12

DISCUSSION

To probe genuine opinions from just legitimate users, proper filtering of Twitter accounts is necessary. The ideal ground truth would be a representative sample of Twitter users who are also voters. We evaluate multiple methods that could be used to obtain such an ideal sample. This article introduces a novel method based on the *#ivoted* hashtag and corresponding pictures to find self-identified voters. Our analyses confirmed that manually-identified voters are least likely to display suspicious behaviors. We also evaluated the performance of two more efficient and lower-cost variants of this method. Using just the *#ivoted* hashtag (*Deb et al., 2019*) is simplest and cheapest, but results based on this sampling method are still subject to some bias along all the dimensions studied. The alternative automated approach is also efficient in that no human resources are required for manual annotation, at the cost of losing some precision and recall. Our systematic comparison reveals that this sampling method is as reliable as the manual method: the `Voter_auto` group is almost identical to `Voter` along all of the dimensions studied.

These methods have some limitations. First, unlike survey-based methods, our approaches cannot provide accurate demographic information about the voters. Second, only voters who are willing to reveal their voting actions online can be identified. This could introduce biases based on, say, age and privacy preferences. Third, the number of identified voters could be limited due to the strict criteria, for example, in cases when there are no hashtags widely adopted to self-identify voters. The resulting smaller size of the sample might negatively affect its statistical power. Finally, the proposed methods could be exploited by malicious actors, *e.g.*, to target voter suppression efforts.

Nevertheless, we believe our approaches have their merit. The manual and automated voter identification methods are efficient and inexpensive compared to traditional ones. They can be applied to future elections in the U.S. and other countries with similar social norms around voting. The methods can be applied to other platforms like Instagram and Facebook where people share photos of themselves wearing "I Voted" stickers.

We further explore the biases in a naive account sampling method based on tweet volume. Compared with the identified voters, accounts in `ByTweet` are hyperactive: they tend to flood the platform with political content and aggressively expand their social

networks while holding a right-leaning political viewpoint. When it comes to questionable behaviors, `ByTweet` accounts tend to show more bot activity and suspicious deletions as well as share more low-credibility information. These findings suggest that the Twitter stream is dominated by these hyperactive accounts, who introduce political bias and inauthentic signals. For researchers, the naive sampling approach could fundamentally bias their studies in the context of political elections. Normal Twitter users could also be disproportionately influenced *via* direct exposure (*Shao et al., 2018*), indirectly *via* social engagement metrics (*Avram et al., 2020*), or through platform mechanisms like trending topics and feed ranking algorithms, which use social engagement metrics as signals (*Ciampaglia et al., 2018*).

Our analyses of the `ByAccount` group suggest that randomly sampling accounts with equal weights could be another effective way to reduce bias and inauthentic signals. Accounts in the `ByAccount` group are similar to likely voters and only exhibit mild suspicious behaviors, although they do have a high suspension rate. For high quality data, one may consider applying bot and troll detection algorithms in the literature to identify and remove likely inauthentic accounts from such a sample (*Sayyadiharikandeh et al., 2020*; *Addawood et al., 2019*).

CONCLUSION

In this article, we propose a method to identify voters on Twitter and systematically compare their behaviors with different samples of Twitter accounts in the context of U.S. midterm elections. We found that compared to likely voters and randomly selected accounts, accounts sampled from the Twitter stream tend to be hyperactive and markedly more conservative. They also exhibit various suspicious behaviors, are involved with more automated activities, and share more links to low-credibility sources. Researchers and journalists who analyze political issues using social media data must be mindful of these pitfalls of sampling biases.

ACKNOWLEDGEMENTS

The authors thank Y.-Y. Ahn for useful comments; D. Pacheco and W. Chen for assistance. We are also grateful for the helpful comments from three anonymous referees.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Knight Foundation and Craig Newmark Philanthropies. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Knight Foundation and Craig Newmark Philanthropies.

Competing Interests

Filippo Menczer is an Academic Editor for PeerJ.

Author Contributions

- Kai-Cheng Yang conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Pik-Mai Hui conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Filippo Menczer conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available at Zenodo: Kai-Cheng Yang, Pik-Mai Hui, & Filippo Menczer. (2022). Dataset for paper: How Twitter Data Sampling Biases U.S. Voter Behavior Characterizations [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6547792.

REFERENCES

Addawood A, Badawy A, Lerman K, Ferrara E. 2019. Linguistic cues to deception: identifying political trolls on social media. In: In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13, 15–25.

Almuhimedi H, Wilson S, Liu B, Sadeh N, Acquisti A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In: *CSCW '13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. 897–908.

Alothali E, Zaki N, Mohamed EA, Alashwal H. 2018. Detecting social bots on Twitter: a literature review. In: *2018 International Conference on Innovations in Information Technology (IIT)*. 175–180.

An J, Kwak H, Ahn Y-Y. 2018. SemAxis: a lightweight framework to characterize domain-specific word semantics beyond sentiment. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2450–2461.

Ausserhofer J, Maireder A. 2013. National politics on Twitter: structures and topics of a networked public sphere. *Information, Communication & Society* **16(3)**:291–314 DOI 10.1080/1369118X.2012.756050.

Avram M, Micallef N, Patil S, Menczer F. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review* **1(5)**:e033 DOI 10.37016/mr-2020-033.

Badawy A, Ferrara E, Lerman K. 2018. Analyzing the digital traces of political manipulation: the 2016 Russian interference Twitter campaign. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Piscataway: IEEE, 258–265.

Bakshy E, Messing S, Adamic L. 2015. Replication data for: exposure to ideologically diverse news and opinion on Facebook. *Harvard Dataverse* DOI 10.7910/DVN/LDJ7MS.

Barberá P. 2016. Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. *Working Paper NYU. Available at http://pablobarbera.com/static/less-is-more.pdf*.

Bessi A, Ferrara E. 2016. Social bots distort the 2016 US presidential election online discussion. *First Monday* **21(11)**:1–14 DOI 10.5210/fm.v21i11.7090.

Bovet A, Makse HA. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* **10(1)**:1–14 DOI 10.1038/s41467-018-07761-2.

Burnap P, Gibson R, Sloan L, Southern R, Williams M. 2016. 140 characters to victory?: using Twitter to predict the uk 2015 general election. *Electoral Studies* **41(2)**:230–233 DOI 10.1016/j.electstud.2015.11.017.

Ceron A, Curini L, Iacus SM, Porro G. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* **16(2)**:340–358 DOI 10.1177/1461444813480466.

Chen W, Pacheco D, Yang K-C, Menczer F. 2021. Neutral bots probe political bias on social media. *Nature Communications* **12(1)**:5580 DOI 10.1038/s41467-021-25738-6.

Ciampaglia GL, Nematzadeh A, Menczer F, Flammini A. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports* **8(1)**:15951 DOI 10.1038/s41598-018-34203-2.

Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A, Menczer F. 2011. Political polarization on Twitter. In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Cota W, Ferreira SC, Pastor-Satorras R, Starnini M. 2019. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science* **8(1)**:35 DOI 10.1140/epjds/s13688-019-0213-9.

Deb A, Luceri L, Badaway A, Ferrara E. 2019. Perils and challenges of social media and election manipulation analysis: the 2018 US midterms. In: *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference*. 237–247.

Diakopoulos NA, Shamma DA. 2010. Characterizing debate performance via aggregated Twitter sentiment. In: *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1195–1198.

DiGrazia J, McKelvey K, Bollen J, Rojas F. 2013. More tweets, more votes: social media as a quantitative indicator of political behavior. *PLOS ONE* **8(11)**:e79449 DOI 10.1371/journal.pone.0079449.

Ferrara E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* **22(8)**:8005 DOI 10.5210/fm.v22i8.8005.

Ferrara E, Chang H, Chen E, Muric G, Patel J. 2020. Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday* **25(11)**:11431 DOI 10.5210/fm.v25i11.11431.

Ferrara E, Varol O, Davis C, Menczer F, Flammini A. 2016. The rise of social bots. *Communications of the ACM* **59(7)**:96–104 DOI 10.1145/2818717.

Gayo-Avello D, Metaxas PT, Mustafaraj E. 2011. Limits of electoral predictions using Twitter. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Gorodnichenko Y, Pham T, Talavera O. 2021. Social media, sentiment and public opinions: evidence from #Brexit and #USElection. *European Economic Review* **136**:103772 DOI 10.1016/j.euroecorev.2021.103772.

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363(6425)**:374–378 DOI 10.1126/science.aau2706.

Guess A, Nagler J, Tucker J. 2019. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5(1)**:eaau4586 DOI 10.1126/sciadv.aau4586.

Hughes A, Jones B, Tyson A, Remy E, Smith A, Kessel Pv, Widjaya R, Wojcik S. 2019. National politics on Twitter: small share of U.S. Adults produce majority of tweets. Pew Research Center. *Available at https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/10/PDL_10.23.19_politics.twitter_FULLREPORT.pdf*.

Hutto CJ, Gilbert E. 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Jungherr A. 2014. The logic of political coverage on Twitter: temporal dynamics and content. *Journal of Communication* **64(2)**:239–259 DOI 10.1111/jcom.12087.

Jungherr A. 2016. Twitter use in election campaigns: a systematic literature review. *Journal of Information Technology & Politics* **13(1)**:72–91 DOI 10.1080/19331681.2015.1132401.

Kozlowski AC, Taddy M, Evans JA. 2019. The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review* **84(5)**:905–949 DOI 10.1177/0003122419877135.

Livne A, Simmons M, Adar E, Adamic L. 2011. The party is over here: structure and content in the 2010 election. In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Barcelona, Catalonia, Spain.

Meeks L. 2018. Tweeted, deleted: theoretical, methodological, and ethical considerations for examining politicians' deleted tweets. *Information, Communication & Society* **21(1)**:1–13 DOI 10.1080/1369118X.2016.1257041.

Mellon J, Prosser C. 2017. Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. *Research & Politics* **4(3)**:2053168017720008 DOI 10.1177/2053168017720008.

Mikolov T, Chen K, Corrado G, Dean J. 2013a. Efficient estimation of word representations in vector space. *ArXiv preprint* DOI 10.48550/arXiv.1301.3781.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013b. Distributed representations of words and phrases and their compositionality. In: *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*. 3111–3119.

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA. 2010. From tweets to polls: linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*. Washington, D.C., USA.

Pennycook G, Rand DG. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences of the United States of America* **116(7)**:2521–2526 DOI 10.1073/pnas.1806781116.

Ramteke J, Shah S, Godhia D, Shaikh A. 2016. Election result prediction using Twitter sentiment analysis. In: In: *2016 International Conference on Inventive Computation Technologies (ICICT)*. Vol. 1. Piscataway: IEEE, 1–5.

Sayyadiharikandeh M, Varol O, Yang K-C, Flammini A, Menczer F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In: *CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*. New York: ACM, 2725–2732.

Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F. 2018. The spread of low-credibility content by social bots. *Nature Communications* **9(1)**:1–9 DOI 10.1038/s41467-018-06930-7.

Small TA. 2011. What the Hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society* **14(6)**:872–895 DOI 10.1080/1369118X.2011.554572.

Stella M, Ferrara E, De Domenico M. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences of the United States of America* **115(49)**:12435–12440 DOI 10.1073/pnas.1803470115.

Stocking G, Sumida N. 2018. *Social media bots draw public's attention and concern*. Washington, D.C.: Pew Research Center.

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vosoughi S, Roy D, Aral S. 2018. The spread of true and false news online. *Science* **359(6380)**:1146–1151 DOI 10.1126/science.aap9559.

Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. 2012. A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle. In: *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea: Association for Computational Linguistics, 115–120.

Wojcik S, Hughes A. 2019. Sizing up Twitter users. Pew Research Center. *Available at https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/04/twitter_opinions_4_18_final_clean.pdf*.

Yang K-C, Hui P-M, Menczer F. 2019a. Bot electioneering volume: visualizing social bot activity during elections. In: *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference*. 214–217.

Yang K-C, Varol O, Davis CA, Ferrara E, Flammini A, Menczer F. 2019b. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* **1(1)**:48–61 DOI 10.1002/hbe2.115.

Yang K-C, Varol O, Hui P-M, Menczer F. 2020. Scalable and generalizable social bot detection through data selection. *Proceedings of the AAAI Conference on Artificial Intelligence* **34(1)**:1096–1103 DOI 10.1609/aaai.v34i01.5460.

Zannettou S, Caulfield T, De Cristofaro E, Sirivianos M, Stringhini G, Blackburn J. 2019a. Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the web. In: *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference*. 218–226.

Zannettou S, Caulfield T, Setzer W, Sirivianos M, Stringhini G, Blackburn J. 2019b. Who let the trolls out? Towards understanding state-sponsored trolls. In: *WebSci '19: Proceedings of the 10th ACM Conference on Web Science*. 353–362.

Zhou L, Wang W, Chen K. 2016. Tweet properly: analyzing deleted tweets to understand and identify regrettable ones. In: *WWW '16: Proceedings of the 25th International Conference on World Wide Web*. 603–612.

PeerJ Computer Science

RuSentiTweet: a sentiment analysis dataset of general domain tweets in Russian

Sergey Smetanin

Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, Russia

ABSTRACT

The Russian language is still not as well-resourced as English, especially in the field of sentiment analysis of Twitter content. Though several sentiment analysis datasets of tweets in Russia exist, they all are either automatically annotated or manually annotated by one annotator. Thus, there is no inter-annotator agreement, or annotation may be focused on a specific domain. In this article, we present RuSentiTweet, a new sentiment analysis dataset of general domain tweets in Russian. RuSentiTweet is currently the largest in its class for Russian, with 13,392 tweets manually annotated with moderate inter-rater agreement into five classes: Positive, Neutral, Negative, Speech Act, and Skip. As a source of data, we used Twitter Stream Grab, a historical collection of tweets obtained from the general Twitter API stream, which provides a 1% sample of the public tweets. Additionally, we released a RuBERT-based sentiment classification model that achieved $F_1 = 0.6594$ on the test subset.

Subjects Computational Linguistics, Data Mining and Machine Learning, Natural Language and Speech, Network Science and Online Social Networks, Social Computing
Keywords Sentiment dataset, Sentiment analysis, Russian

INTRODUCTION

Recently, Twitter has been established as a major research platform, utilized in more than ten thousand research articles over the past ten years. Sentiment analysis has proven to be one of the major research areas (*Antonakaki, Fragopoulou & Ioannidis, 2021*). As expected, there is an interest in the sentiment analysis of the Russian-speaking segment of Twitter, not only for training machine learning (ML) models (*Kotelnikova, 2020*; *Araslanov, Komotskiy & Agbozo, 2020*; *Kanev et al., 2022*), but also for applied research—such as studying migration issues (*Borodkina & Sibirev, 2019*), measuring reactions to different events (*Kirilenko & Stepchenkova, 2017*; *Kausar, Soosaimanickam & Nasar, 2021*), and monitoring public sentiment (*Chizhik, 2016*; *Smetanin, 2017*). However, despite the fact that there are several datasets of tweets in Russian (*Smetanin, 2020a*), they are either annotated automatically (*e.g.*, RuTweetCorp by *Rubtsova (2013)*) or annotated only by one annotator; thus, there is no inter-annotator agreement (*e.g.*, Twitter Sentiment for 15 European Languages by *Mozeticar & Smailović (2016)*), or focused on a specific domain (*e.g.*, SentiRuEval-2015 by *Loukachevitch et al. (2015)*). Thus, this research community lacks a general domain sentiment dataset of tweets in Russian that is annotated manually with reported inter-rater agreement score.

Submitted 12 May 2022
Accepted 22 June 2022
Published 19 July 2022

Corresponding author
Sergey Smetanin,
sismetanin@gmail.com

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj-cs.1039**

Copyright
2022 Smetanin

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

In this article, we present RuSentiTweet, a new sentiment analysis dataset of 13,392 general domain tweets in Russian. RuSentiTweet was annotated manually using RuSentiment guidelines (*Rogers et al., 2018*) into five classes (*Positive*, *Neutral*, *Negative*, *Speech Act*, and *Skip*) with moderate inter-rater agreement. The practical and academic contribution of this study is threefold. Firstly, we reviewed existing public sentiment dataset of tweets in Russian. Secondly, we filled the data gap and introduced RuSentiTweet, the only dataset of general domain tweets with manual annotation for the Russian language. Lastly, we trained several ML models to provide further research with a strong baseline.

The rest of the article is organized as follows. In "Related Work", we review related research, identify existing public sentiment datasets of tweets in Russian, and confirm the importance of a new dataset of general domain tweets in Russian. In "Sentiment Dataset", we describe the creation of RuSentiTweet. In "Sentiment Classification Baseline", we document the training of several ML models to provide the research community with public baselines. In "Conclusion", we present conclusions from this study.

RELATED WORK

As of 2022, Russian was the eighth most widely-spoken language worldwide, with a total number of 258.2 million speakers (*Szmigiera, 2022*). Yet as reported in the preliminary results of the All-Russian Census 2020 (*Rosstat, 2022*), only about 147 million people permanently live in Russia. In addition to Russia, where Russian is the official language, it is also widely spoken in a number of other countries that were part of the USSR. According to various sources (*Arefiev, 2013*; *Lopatin & Ulukhanov, 2017*), there are from 52 to 94 million native speakers of the Russian language in these countries. A large number of Russian speakers also live in other countries such as those in Europe, the USA, Canada, Israel, and others (*Lopatin & Ulukhanov, 2017*). Given the significant Russian-speaking population and the ever-growing level of Internet penetration, texts published by Russian-speaking users on social networks are attracting more and more attention from researchers. As a result, every year new works appear both in the classical analysis of the sentiment of Russian-language content (*e.g.*, *Araslanov, Komotskiy & Agbozo, 2020*; *Kanev et al., 2022*; *Kausar, Soosaimanickam & Nasar, 2021*) and in related areas, such as the identification of emotions (*e.g.*, *Babii, Kazyulina & Malafeev, 2020*; *Kazyulina, Babii & Malafeev, 2020*; *Babii, Kazyulina & Malafeev, 2021*), toxicity and hate speech detection (*e.g.*, *Zueva, Kabirova & Kalaidin, 2020*; *Pronoza et al., 2021*; *Smetanin & Komarov, 2021b*), and inappropriate language identification (*e.g.*, *Babakov et al., 2021*; *Babakov, Logacheva & Panchenko, 2022*).

However, the Russian language is not as well-resourced as the English language (*Besacier et al., 2014*), especially in the field of sentiment analysis (*Smetanin & Komarov, 2021a*), so the data options for researchers are quite limited. In our previous study (*Smetanin, 2020a*), we identified 14 publicly available sentiment analysis datasets of Russian texts. In said study, we considered only those datasets that can be accessed *via* instructions from their original papers or official websites. Following this strategy, we omitted several existing datasets—such as ROMIP datasets (*Chetviorkin, Braslavskiy &*

Table 1 Sentiment analysis datasets of Russian language texts. More detailed description of each datasetcan be found in *Smetanin (2020a)*, *Smetanin & Komarov (2021a)*, *Kotelnikov (2021)*, as well as in original papers (if published). For datasets that contain several subsets from different data sources, we indicated only those subsets that are made from tweets.

Dataset	Data source	Domain	Annotation	Classes	Size	Link
Twitter Sentiment for 15 European Languages (*Mozeticar & Smailović, 2016*)	Twitter	General	Manual	3	107,773	Project page
SemEval-2016 Task 5 Russian (*Pontiki et al., 2016*)	Twitter	Restaurants	Manual	3	405	Project page
SentiRuEval-2016 (*Lukashevich & Rubtsova, 2016*)	Twitter	Telecom and banks	Manual	3	23,595	Project page
SentiRuEval-2015 (*Loukachevitch et al., 2015*)	Twitter	Telecom and banks	Manual	4	16,318	Project page
RuTweetCorp (*Rubtsova, 2013*)	Twitter	General	Automatic	3	334,836	Project page
Kaggle Russian_twitter_sentiment	Twitter	n/a	n/a	2	226,832	Kaggle page

Loukachevich, 2013; *Chetvirokin & Loukachevitch, 2013*)—because we were unable to obtain access to them. Among these 14 datasets, only six datasets were constructed based on Twitter content, so we selected them for further detailed analysis. Additionally, we analysed the most recent review of sentiment analysis datasets of Russian texts by *Kotelnikov (2021)* but did not find any new Twitter datasets for consideration.

As can be seen from Table 1, RuTweetCorp (*Rubtsova, 2013*) is the largest sentiment analysis dataset of general domain tweets in Russian, but it was automatically annotated based on the strategy proposed by *Read (2005)*: each tweet was assigned with the sentiment class based on the emoticons it contains. As a consequence, even a simple rule-based approach based on the presence of the '(' character can achieve $F_1 = 97.39\%$ in the binary (*Positive* and *Negative* classes) classification task (*Smetanin & Komarov, 2021a*). SemEval-2016 Task 5 Russian (*Pontiki et al., 2016*), SentiRuEval-2016 (*Lukashevich & Rubtsova, 2016*) and SentiRuEval-2015 (*Loukachevitch et al., 2015*) are manually annotated and widely used datasets, but they are all tied to a specific domain such as restaurants, automobiles, telecommunication companies, or banks. Twitter Sentiment for 15 European Languages (*Mozeticar & Smailović, 2016*) is a sentiment analysis dataset with manual annotation, but only one annotator was engaged for Russian-language tweets; thus, there is no inter-annotator agreement. The Kaggle dataset did not report data collection and annotation procedure. Thus, there is a lack of general domain sentiment dataset of tweets in Russian that is annotated manually with reported inter-rater agreement score.

SENTIMENT DATASET

Data collection

For a data source of tweets in Russian, we decided to use the Twitter Stream Grab (https://archive.org/details/twitterstream), a publicly available historical collection of JSON grabbed from the general Twitter "Spritzer" API stream. According to Twitter, this API provides a 1% sample of the complete public tweets and is not tied to a specific topic, so we considered it as a good source of general domain tweets. Additionally, several studies (*Wang, Callan & Zheng, 2015*; *Leetaru, 2019*) performed independent validation of the representativeness of this stream. Since the Twitter Stream Grab consists of tweets in

different languages, our first step was to remove tweets written in non-Russian languages. Each tweet from this data source already contained information about the language of the text automatically detected[1] by Twitter, so the language filtering procedure was fairly straightforward.

We downloaded the Twitter Stream Grab for 12 months from January 2020 to December 2020[2]. The main motivation for choosing an entire year as the interval was to cover all months of the year to minimize the effect of seasonality. Previous research has shown that there are daily (*Larsen et al., 2015*; *Prata et al., 2016*), weekly (*Ten Thij, Bhulai & Kampstra, 2014*; *Dzogang, Lightman & Cristianini, 2017b*), and seasonal (*Dzogang et al., 2017a*) patterns of sentiment or emotion expression on Twitter. Also, it has been found (*Baylis et al., 2018*; *Baylis, 2020*) that expressed sentiment correlates with weather, which also tends to depend on the season. After excluding retweets and filtering by language, we obtained ~4.5M tweets in Russian. Since manual labelling of such a volume of tweets is costly and extremely time-consuming, we randomly selected 15,000 tweets for further annotation (tweets evenly distributed over the selected months).

[1] Assessing the quality of a given algorithm lies outside the scope of this study. Initial research in this direction has already been done in other studies; for example, *Pavliy and Lewis (2016)* compared the quality of Twitter's language detection algorithm and Google's Compact Language Detector on Ukrainian and Russian tweets. The authors found that Twitter's algorithm correctly detects 92% of texts in Russian and has higher accuracy than Google's Compact Language Detector.

[2] At the time of this writing, all months for 2021 were not available.

Data annotation

Guidelines

As per recommendations outlined in our previous study (*Smetanin, 2020a*), we decided to use RuSentiment (*Rogers et al., 2018*) annotation guidelines (https://github.com/text-machine-lab/rusentiment/tree/master/Guidelines). To the best of our knowledge, this is the only set of publicly available sentiment annotation guidelines designed for the Russian language. The guidelines are described in detail in the original RuSentiment paper, so this section provides only key summary.

The annotation guidelines cover both implicit and explicit forms of expressions for external attitude (evaluation) and the internal emotional state (mood). The guidelines cover five sentiment classes.

- *Negative* represents both explicit and implicit negative sentiment or attitude towards something.
- *Neutral* represents texts that simply describe some situation in a neutral, matter-of-fact way and have no clear positive or negative sentiment. This class also includes commercial information, factual questions, objective descriptions, and summaries.
- *Positive* represents both explicit and implicit positive sentiment or attitude towards something.
- *Speech Act* represents texts that perform the functions of various speech acts—such as greeting someone, congratulating someone, and expressing gratitude for something. Although these texts also represent a positive sentiment, they are treated as a separate subcategory because they can also be performed under social pressure or out of a feeling of obligation (*Rogers et al., 2018*).
- *Skip* represents noisy and unclear sentiment or attitude towards something—such as when the original meaning is impossible to ascertain without additional context, the

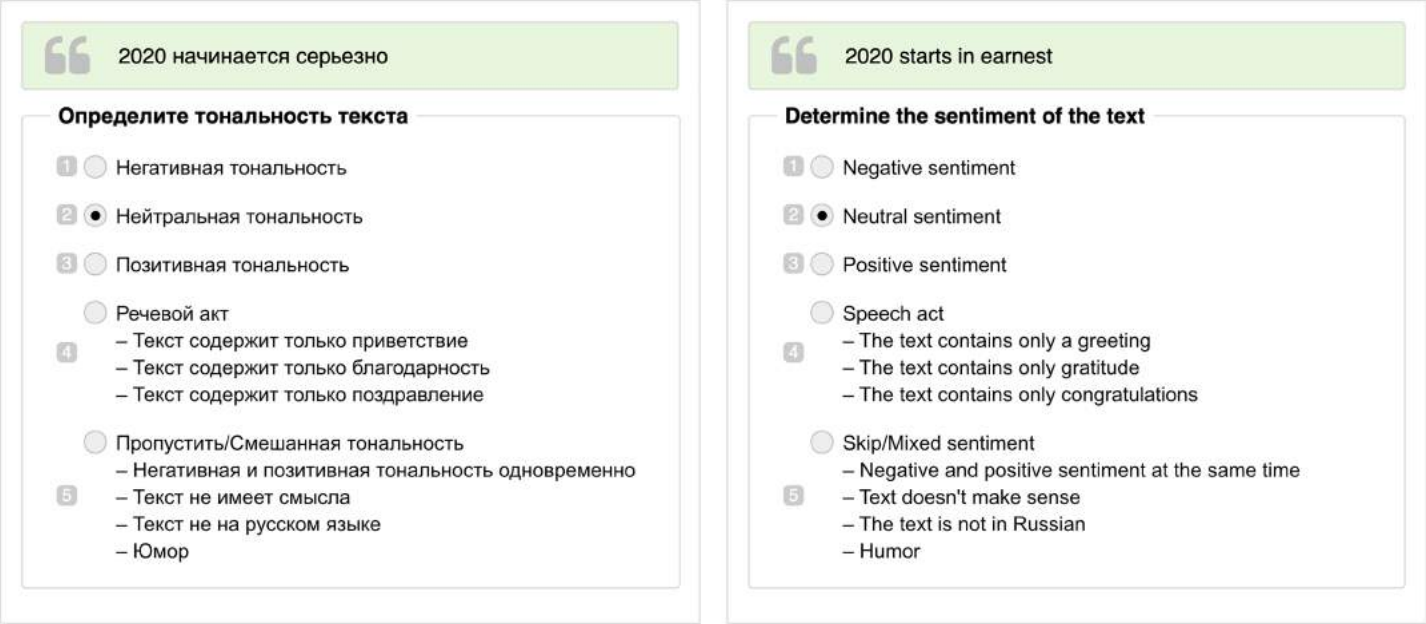


Figure 1 An example of user interface for annotators in Yandex.Toloka in Russian (on the left) and its translation in English (on the right). The green block with quotation marks contains the text of the tweet. Under the block with the text, there are numbered sentiment classes, where 1 is *Negative*, 2 is *Neutral*, 3 is *Positive*, 4 is *Speech Act*, and 5 is *Skip*. Numbers are used as hotkeys during annotation.
Full-size DOI: 10.7717/peerj-cs.1039/fig-1

sentiment of the texts as a whole is not entirely clear, the text is not in Russian, or the text contains jokes.

Text with irony was annotated with the dominant sentiment, commonly negative. Hashtags were treated as information units similar to basic words or phrases. Emoticons were not treated as the only sentiment labels but were analysed in combination with the whole text to identify dominant sentiment.

Crowdsourcing platform

The annotation was performed *via* Yandex.Toloka (https://toloka.ai/), a Russian crowdsourcing platform with a high share of Russian speaking workers. Yandex.Toloka is widely used in the studies on Russian-language content, such as for annotation of semantic change (*Rodina & Kutuzov, 2020*), question answering (*Korablinov & Braslavski, 2020*), and toxic comments (*Smetanin, 2020b*). A depiction of the Yandex.Toloka user interface can be found in Fig. 1. We required annotators to pass training before starting annotation. During the annotation of the dataset, their work was continuously evaluated through honeypots. As training samples and control pairs, we selected texts from RuSentiment, which was annotated using the same guidelines. The threshold was 60% correctly annotated samples for training and 80% samples for honeypots. We selected only Russian speaking annotators who passed an internal exam (https://toloka.ai/ru/docs/guide/concepts/filters.html) on language knowledge.

Aggregation

Following the RuSentiment aggregation strategy, a tweet was deemed to belong to a class if at least two out of the three annotators attributed it to that class. In case all three annotators disagreed, the tweet was removed from the dataset as extremely noisy and unclear (see examples in Table A1). Out of the initially selected 15,000 tweets, 1,608 tweets received all three different annotations, so we excluded these tweets from the final dataset. Thus, the final dataset consists of 13,392 tweets with the following class distribution: 3,298 (24.62%) *Negative* tweets, 5,341 (39.88%) *Neutral* tweets, 2,414 (18.02%) *Positive* tweets, 1,843 (13.76%) *Skip* tweets, and 496 (3.70%) *Speech Act* tweets. We split our dataset into training (80%) and test subset (20%) using stratified random sampling by class labels.

Inter-annotator agreement

For measuring the inter-annotator agreement, we calculated the Krippendorff's α coefficient (*Krippendorff, 1980*) because it applies to any number of annotators and categories, as well as to missing or incomplete data (*Krippendorff, 2004*). For most inter-annotator agreement indices, including Krippendorff's α, it is commonly suggested that a cutoff threshold value of 0.8 is a marker of good reliability, with a range of 0.667 to 0.8 allowing for tentative conclusions and values below 0.667 indicating poor agreement (*Beckler et al., 2018*). However, in the systematic review of crowd-sourced annotation in social computing, *Salminen et al. (2018)* reported that agreement scores in social computing studies are not high, averaging at around 0.60 for both Kappa and Alpha metrics, which is lower than typical threshold values. The authors highlighted that the nature of annotation in social computing tends to be more subjective rather than objective, and the more subjective the task, the worse the agreement, regardless of the metric used. Though it is important to report inter-rater agreement scores, there are suggestions that the results can be misleading in social computing (*Hillaire, 2021*). In fact, low agreement in this case does not necessarily mean the opinions of annotators are incorrect; it may simply indicate that they have different opinions (*Salminen et al., 2018*; *Hillaire et al., 2021*). Sentiment annotation, by nature, is a subjective task because the annotator must subjectively (with some guidelines) identify sentiment and emotions expressed by the author and not just objectively analyse narrated events or situations: we can expect annotators to have different subjective understanding of emotion expressed in a particular text. Thus, considering that in the field of social computing science the mean score is 0.60 (*Salminen et al., 2018*), we followed the same approach as *Hillaire (2021)* and adopted the less conservative interpretation of inter-rater agreement by *Landis & Koch (1977)*, which suggests the following interpretations.

- Scores from 0.0 to 0.2 indicate a slight agreement.
- Scores from 0.21 to 0.40 indicate a fair agreement.
- Scores from 0.41 to 0.60 indicate a moderate agreement.
- Scores from 0.61 to 0.80 indicate a substantial agreement.
- Scores from 0.81 to 1.0 indicate almost perfect or perfect agreement.

Table 2 Distance between classes for interval Krippendorff's α, where 0 means that classes are the same, 1 means that classes are close to each other, and 2 means that classes a far away from each other.

Class	Negative	Neutral	Positive	Speech	Skip
Negative	0	1	2	2	1
Neutral	1	0	1	1	1
Positive	2	1	0	0	1
Speech	2	1	0	0	1
Skip	1	1	1	1	0

Note:
Positive and *Speech* classes have zero distance between them; they both represent positive sentiment as per RuSentiment guidelines.

We calculated Krippendorff's α with binary distance (*e.g.*, all classes have similar distance between each other) using the NLTK library (*Bird, Klein & Loper, 2009*) and obtained the score of 0.5048 for binary distance, which can be interpreted as a moderate agreement between annotators. We considered this level of agreement as satisfactory for our case, since other five-class sentiment datasets also reported this or even lower level of agreement, such as Blog Track at TREC 2008 ($\alpha = 0.4219$, five classes) (*Bermingham & Smeaton, 2009*), LINIS Crowd ($\alpha = 0.541$, five classes) (*Koltsova, Alexeeva & Kolcov, 2016*), RuSentiment (Fleiss' kappa of 0.58, five classes) (*Rogers et al., 2018*), sentiment@USNavy ($\alpha = 0.592$, four classes) (*Fiok et al., 2021*), and NaijaSenti (Fleiss kappa of (0.434, 0.555), five classes) (*Muhammad et al., 2022*). Additionally, we calculated Krippendorff's α with interval distance that takes into account distance between classes: for example, *Neutral* and *Positive* classes are closer to each other than *Negative* and *Positive* classes. The distance matrix is presented in Table 2. The Krippendorff's α coefficient for interval distance was 0.5601, which can also be interpreted as slightly higher but still moderate agreement.

Explanatory analysis

The average text length is 59.36 characters for all text, 67.52 for *Negative*, 59.29 for *Neutral*, 57.85 for *Positive*, 42.71 for *Speech*, and 51.41 for *Skip*. As can be seen from Fig. 2, the frequency of occurrence of texts from a pair of characters in the dataset is extremely low, but with an increase in the number of characters, rapid growth begins. The frequency peak is reached when the text length is from 20 to 40 characters, and then the frequency gradually begins to decrease. Interestingly, for some classes, there is a moderate Pearson's correlation between the length of the text and the proportion of texts with this class relative to all texts. The *Negative* class has a moderate positive correlation ($\rho = 0.68$, $p < 0.01$) with the length of text, whereas *Speech* ($\rho = -0.52$, $p < 0.01$) and *Skip* ($\rho = -0.62$, $p < 0.01$) classes have moderate negative correlation. At the same time, *Neutral* ($\rho = -0.03$, $p = 0.70$) and *Positive* ($\rho = -0.04$, $p = 0.62$) classes do not have statistically significant correlation. The most common unigrams, bigrams, and emojis can be found in Table 3.

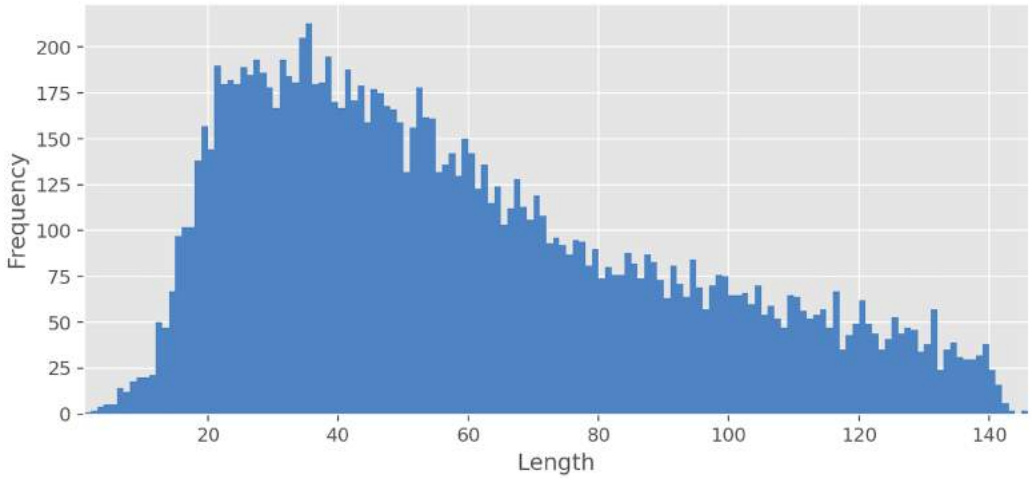


Figure 2 **Texts length distribution.** Full-size DOI: 10.7717/peerj-cs.1039/fig-2

Table 3 **Most common unigrams, bigrams, and emojis without stop words, punctuation, and numbers.** Stop words were removed using NLTK (*Bird, Klein & Loper, 2009*). Most unigrams and bigrams can have several English translations depending on the context. The table provides only one translation option.

Unigram			Bigram			Emoji	
Item		Count	Item		Count	Item	Count
Russian	English		Russian	English			
это	it	1,117	доброе утро	good morning	39		443
просто	simply	355	спокойной ночи	good night	26		313
спасибо	thanks	306	спасибо большое	thanks a lot	24		246
хочу	want	249	самом деле	actually	23		240
ещё	yet	223	это просто	it's simple	23		120
почему	why	209	опубликовано фото	published photo	18		119
очень	very	205	сих пор	so far	17		118
всё	all	204	руб г	rub g	16		113
блять	fuck	184	днем рождения	birthday	15		104
вообще	generally	174	все ещё	still	13		100

As mentioned in "Related Work", one of the key limitations of RuTweetCorp (*Rubtsova, 2013*)—the biggest automatically annotated dataset of tweets in Russian—is that *Positive* and *Negative* tweets in it can be easily separated with F_1 = 97.39% by a simple rule-based approach based on the presence of the '(' character. We decided to check that this limitation does not apply to RuSentiTweet. We applied this simple rule-based approach to *Positive* and *Negative* tweets from RuSentiTweet and got F_1 = 0.3450 (*i.e.*, approximately the same result as in the case of a random classification), thereby confirming that RuTweetCorp's limitation does not apply to RuSentiTweet.

Table 4 Five-class sentiment classification on RuSentiTweet.

Model	Precision	Recall	F_1^{macro}	$F_1^{weighted}$
RuBERT	0.6793	0.6449	0.6594	0.6675
MNB	0.5867	0.5021	0.5216	0.5189

SENTIMENT CLASSIFICATION BASELINE

Model selection

As was demonstrated in our recent study (*Smetanin & Komarov, 2021a*), sentiment analysis of the Russian language text based on the language models tends to outperform rule-based and basic ML-based approaches in terms of classification quality. This statement was also supported by other studies (*Golubev & Loukachevitch, 2020*; *Kotelnikova, 2020*; *Konstantinov, Moshkin & Yarushkina, 2021*). Based on the mentioned papers, we decided to fine-tune RuBERT (*Kuratov & Arkhipov, 2019*), a version of BERT (*Devlin et al., 2019*) trained on the Russian part of Wikipedia and Russian news. Over the past few years, this model has been actively used in sentiment analysis studies on the Russian language and constantly demonstrated strong or even new state-of-the-art (SOTA) results (*Golubev & Loukachevitch, 2020*; *Kotelnikova, 2020*; *Konstantinov, Moshkin & Yarushkina, 2021*; *Smetanin & Komarov, 2021a*). For comparison, we also decided to train more a classical ML classifier for sentiment analysis task: Multinomial Naive Bayes (MNB). We used the MNB implementation (https://github.com/sismetanin/sentiment-analysis-of-tweets-in-russian) from our previous paper (*Smetanin & Komarov, 2019*).

Results

During the training stage for RuBERT, we relied on the approach used in *Smetanin & Komarov (2021a)*. Fine-tuning was performed using the Transformers library (*Wolf et al., 2020*) on 1 Tesla V100 SXM2 32GB GPU with the following parameters: four train epochs, 128 max sequence length, 32 batch size, and a learning rate of 5e−5. Since our goal was to provide a baseline classification model and not the most efficient one, we did not search for the most efficient training parameters. We repeated each experiment 3 times and reported mean values of the measurements. For MNB, we used the same parameters as in our previous paper (*Smetanin & Komarov, 2019*): combination of unigrams and bigrams, TF-IDF vectorizer, and an alpha of 0.01.

According to the results presented in Table 4, RuBERT outperformed MNB, as expected, and showed the best classification scores. The classification results obtained on RuSentiTweet are slightly lower but still comparable with the results obtained in other studies on RuSentiment (see Table 5): RuBERT achieved $F_1^{weighted} = 0.7263$ on RuSentiment (*Kuratov & Arkhipov, 2019*), whereas on our dataset this model showed $F_1^{weighted} = 0.6675$. The difference in the results could be caused by the size of the dataset because RuSentiment is more than two times bigger. The classification metrics of five-class sentiment analysis approaches on other datasets in other languages can be found in

Table 5 Five-class sentiment classification studies.

Study	Dataset	Model	Classification metrics				
			Accuracy	Precision	Recall	F_1^{macro}	$F_1^{weighted}$
Muhammad et al. (2022)	NaijaSenti	XLM-R-base+LAFT	n/a	n/a	n/a	n/a	0.795
Muhammad et al. (2022)	NaijaSenti	M-BERT+LAFT	n/a	n/a	n/a	n/a	0.7700
Fiok et al. (2021)	sentiment@USNavy	BART large + CNN	n/a	n/a	n/a	0.596	n/a
Smetanin & Komarov (2021a)	RuSentiment	M-BERT-Base	n/a	0.6722	0.6907	0.6794	0.7244
Smetanin & Komarov (2021a)	RuSentiment	RuBERT	n/a	0.7089	0.7362	0.7203	0.7571
Smetanin & Komarov (2021a)	RuSentiment	M-USE-CNN	n/a	0.6571	0.6708	0.6627	0.7105
Smetanin & Komarov (2021a)	RuSentiment	M-USE-Trans	n/a	0.6821	0.6982	0.6860	0.7342
Jamadi Khiabani, Basiri & Rastegari (2020)	TripAdvisor	Dempster–Shafer-based model	0.79	0.5	0.47	0.49	n/a
Jamadi Khiabani, Basiri & Rastegari (2020)	CitySearch	Dempster–Shafer-based model	0.79	0.48	0.48	0.48	n/a
Kuratov & Arkhipov (2019)	RuSentiment	Multilingual BERT	n/a	n/a	n/a	n/a	0.7082
Kuratov & Arkhipov (2019)	RuSentiment	RuBERT	n/a	n/a	n/a	n/a	0.7263
Baymurzina, Kuznetsov & Burtsev (2019)	RuSentiment	SWCNN + fastText Twitter	n/a	n/a	n/a	n/a	0.7850
Baymurzina, Kuznetsov & Burtsev (2019)	RuSentiment	BiGRU + ELMo Wiki	n/a	n/a	n/a	n/a	0.6947
Tripto & Ali (2018)	YouTube	LSTM	0.5424	n/a	n/a	0.5320	n/a
Li et al. (2018)	Twitter	Logistic Regression	0.6899	0.6053	0.6899	0.6354	n/a
Ahmadi et al. (2017)	SST-5	RNTN	0.41	n/a	n/a	0.32	n/a
Buntoro, Adji & Purnamasari (2016)	Twitter	Naïve Bayes	0.7177	0.716	0.718	n/a	n/a
Aly & Atiya (2013)	LABR	SVM	0.503	n/a	n/a	n/a	0.491
Chetvirokin & Loukachevitch (2013)	ROMIP-2012 (Movies)	n/a	0.407	n/a	n/a	0.377	n/a
Blinov, Kotelnikov & Pestov (2013)	ROMIP-2012 (Books)	SVM	0.481	0.339	0.496	0.402	n/a
Chetvirokin & Loukachevitch (2013)	ROMIP-2012 (Cameras)	n/a	0.480	n/a	n/a	0.336	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Movies)	SVM	0.599	n/a	n/a	0.286	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Books)	SVM	0.622	n/a	n/a	0.291	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Cameras)	SVM	0.626	n/a	n/a	0.342	n/a

Note:
We selected only those studies, which consideredfive sentiment classes and reported at least one of the following classification measures: Precision, Recall, macro *F1*, weighted *F1*. Among all datasets, only ROMIP (*Chetviorkin, Braslavskiy & Loukachevich, 2013*; *Chetvirokin & Loukachevitch, 2013*) and RuSentiment (*Rogers et al., 2018*) datasets are in Russian.

Table 5. Although direct comparison for different datasets and languages may not be entirely correct, we can see that at least the magnitude of order of our approach corresponds with the average score for five-class classification.

We made our RuBERT-based model publicly available (https://huggingface.co/sismetanin/rubert-rusentitweet) to the research community.

Error analysis

Considering that RuBERT clearly outperformed MNB, we performed error analysis only for RuBERT. As can be seen from confusion matrix for RuBERT (see Fig. 3), the *Skip* class was one of the most scarcely classified classes since it initially consisted of barely

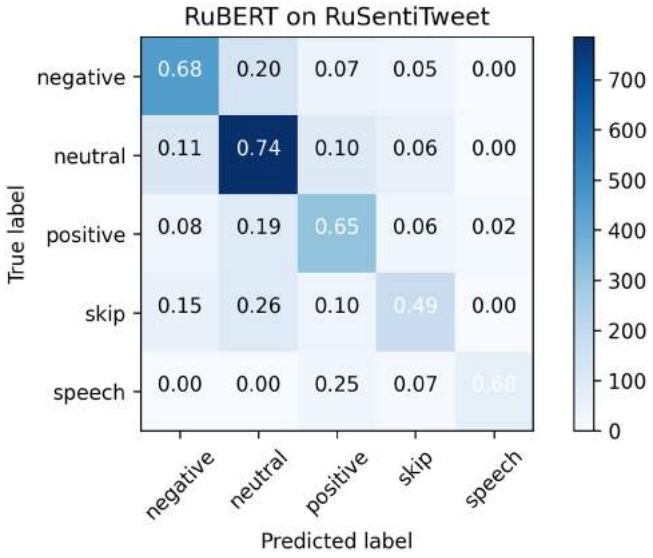


Figure 3 **Confusion matrix for RuSentiTweet.** Full-size DOI: 10.7717/peerj-cs.1039/fig-3

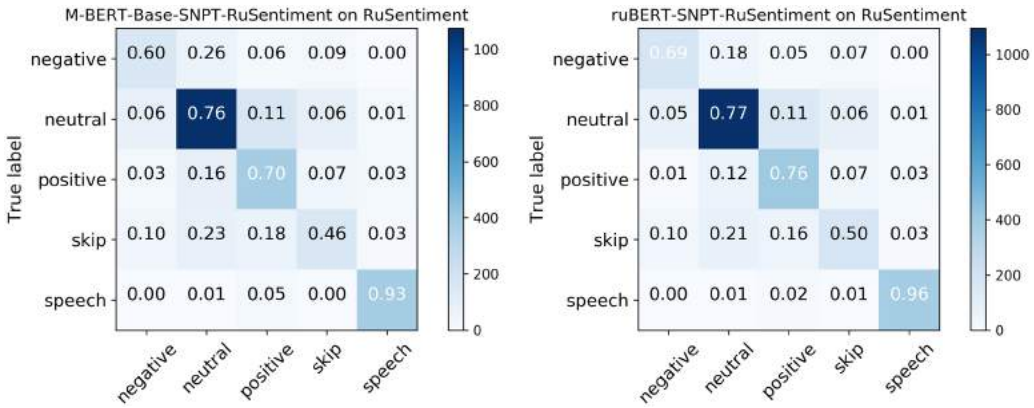


Figure 4 **Confusion matrix for RuSentiment were created using molders from *Smetanin & Komarov (2021a)*.** Full-size DOI: 10.7717/peerj-cs.1039/fig-4

interpretable and noisy tweets. The *Speech Acts* class was clearly distinguished from *Negative* and *Neutral* classes because it consists of a well-defined group of speech constructs, but it was commonly misclassified as *Positive* because it also represents positive sentiment. Predictably, the *Neutral* class was commonly misclassified as *Positive* or *Negative* class because neutral sentiment is logically located between positive and negative sentiment. As was highlighted by *Barnes, Øvrelid & Velldal (2019)*, the issue of neutral sentiment misclassification tends to be a general challenge of non-binary sentiment classification. In general, misclassification errors of our model were quite similar to RuSentiment misclassification errors reported in our previous study (*Smetanin & Komarov, 2021a*) (see Fig. 4), most likely because the same annotation guidelines and models were used. The most noticeable difference was in the recall for the *Speech* class. For RuSentiment, it was much better separated from other classes, with recall in the interval

from 0.88 to 0.96 (*Smetanin & Komarov, 2021a*). We suppose that the reason of such a difference is in the number of texts in this class: RuSentiment contains 3,467 texts of the *Speech*, whereas RuSentiTweet contains only 480 such texts. The examples of misclassified tweets can be found in Table A2.

CONCLUSION

In this article, we present RuSentiTweet, a new general domain sentiment dataset of tweets in Russian with manual annotation. RuSentiTweet includes 13,392 tweets annotated by three annotators with moderate inter-rater agreement into five classes: *Positive*, *Neutral*, *Negative*, *Speech Act*, and *Skip*. Currently, RuSentiTweet is the only dataset of general domain tweets in Russian with manual annotation by more than one annotator and is the largest in its class for Russian. Additionally, we presented a RuBERT-based model trained on RuSentiTweet, which demonstrated $F_1 = 0.6594$ in five-class classification. The code, data, and model were made publicly available to the research community.

Further research might focus on several areas. Firstly, considerably more work must be done to determine the most efficient ML algorithm in terms of classification quality for RuSentiTweet. In particular, it could be interesting to apply explainable sentiment analysis approaches (*e.g.*, *Szczepański et al., 2021*; *Kumar & Raman, 2022*) to allow a deeper understanding of the reasons for misclassification errors on particular texts. Secondly, it would be interesting to measure a subjective well-being index based on historical Russian tweets. Lastly, another possible area of future research would be to perform additional toxicity annotation of negative tweets from RuSentiTweet.

ACKNOWLEDGEMENTS

We thank the authors of the Twitter Stream Grab for collecting public tweets and making them available to the research community. We also thank the authors of RuSentiment for making their annotation guidelines publicly available.

This research was supported in part through computational resources of HPC facilities at HSE University (*Kostenetskiy, Chulkevich & Kozyrev, 2021*).

APPENDIX

Table A1 **Examples of tweets with no agreement between annotators.**

Tweet		Annotation		
Russian	**English**	**Annotator 1**	**Annotator 2**	**Annotator 3**
ю ноу блин	yu know damn it	Skip	Neutral	Negative
@USER доброе утро всем дэдди сегодня	@USER good morning daddies to everyone today	Speech	Positive	Skip
Путешествуем по Уэльсу. не уважать мужчин	Traveling in Wales. disrespect men	Negative	Neutral	Skip
тот факт что в энимал кроссинге так мало прикольных мышиных жителей	the fact that there are so few funny mouse inhabitants in animal crossing	Negative	Positive	Neutral
Кто в нижнем родился, в верхнем не сгодился.	Who was born in the bottom, did not fit in the top.	Negative	Neutral	Skip

Table A2 Examples of tweets classification. All usernames and URLs were replaced with keywords for anonymity purposes.

Tweet		True class	Predicted class
Russian	**English**		
@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME Помедорус	@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME Pomedorus	Skip	Skip
@USERNAME ты не лохушка ЛОЛ я тебе завидую…. у меня травма из за интернета вот я лохушка	@USERNAME you're not a sucker LOL I envy you…. I'm traumatized because of the internet I'm a sucker	Skip	Negative
@USERNAME Котиков Одриосолу Дождь	@USERNAME Cats Odriosolu Rain	Skip	Neutral
@USERNAME Уж лучше твоя грудь	@USERNAME Your breasts are better	Skip	Positive
Как сережки URL	How do you like the earrings URL	Neutral	Neutral
@USERNAME Реквием по мечте	@USERNAME Requiem for a dream	Neutral	Positive
@USERNAME ПОДОЖДИ НУ МНЕ КАЗАЛОСЬ ДА	@USERNAME WAIT I THINK YES	Neutral	Negative
@USERNAME Спокойной ночи и сладких снов	@USERNAME Good night and sweet dreams	Speech	Speech
@USERNAME как дела моя хорошая?? ( )	@USERNAME how are you my dear?? ( )	Speech	Positive
@USERNAME Это классно что у тебя есть эти люди	@USERNAME It's great that you have these people	Positive	Positive
На самом деле я ловлю уруру с этого облачка.	In fact, I catch ururu from this cloud.	Positive	Neutral
@USERNAME Это классно что у тебя есть эти люди	What kind of morons are you, many have a school/work day tomorrow	Negative	Negative
интересный факт: смысла в клипах txt больше, чем в твоей жизни	Interesting fact: there is more sense in txt clips than in your life	Negative	Neutral

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

There was no external funding received for this study.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Sergey Smetanin conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code and dataset are available at GitHub: https://github.com/sismetanin/rusentitweet.

The sentiment classification model is available at HuggingFace: https://huggingface.co/sismetanin/rubert-rusentitweet.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1039#supplemental-information.

REFERENCES

Ahmadi Z, Skowron M, Stier A, Kramer S. 2017. An in-depth experimental comparison of RNTNs and CNNs for sentence modeling. In: *International Conference on Discovery Science*. Springer, 144–152.

Aly M, Atiya A. 2013. LABR: a large scale Arabic book reviews dataset. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Vol. 2. Association for Computational Linguistics, 494–498.

Antonakaki D, Fragopoulou P, Ioannidis S. 2021. A survey of Twitter research: data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications* **164**:114006 DOI 10.1016/j.eswa.2020.114006.

Araslanov E, Komotskiy E, Agbozo E. 2020. Assessing the impact of text preprocessing in sentiment analysis of short social network messages in the Russian language. In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. Piscataway: IEEE, 1–4.

Arefiev A. 2013. Demographic changes are not good for the Russian language. *Demoscope Weekly* 571–572.

Babakov N, Logacheva V, Kozlova O, Semenov N, Panchenko A. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, 26–36.

Babakov N, Logacheva V, Panchenko A. 2022. Beyond plain toxic: detection of inappropriate statements on flammable topics for the Russian language. *Available at https://arxiv.org/abs/2203.02392*.

Babii A, Kazyulina M, Malafeev AY. 2020. Automatic emotion identification in Russian text messages. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2020*. Russian State University for the Humanities, 1002–1010.

Babii A, Kazyulina M, Malafeev A. 2021. FastText-based methods for emotion identification in Russian internet discourse. In: *WebSci '21: 13th ACM Web Science Conference 2021*. 112–119.

Barnes J, Øvrelid L, Velldal E. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, 12–23.

Baylis P. 2020. Temperature and temperament: evidence from Twitter. *Journal of Public Economics* **184**:104161 DOI 10.1016/j.jpubeco.2020.104161.

Baylis P, Obradovich N, Kryvasheyeu Y, Chen H, Coviello L, Moro E, Cebrian M, Fowler JH. 2018. Weather impacts expressed sentiment. *PLOS ONE* **13(4)**:e0195750 DOI 10.1371/journal.pone.0195750.

Baymurzina D, Kuznetsov D, Burtsev M. 2019. Language model embeddings improve sentiment analysis in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. Vol. 18. 53–63.

Beckler DT, Thumser ZC, Schofield JS, Marasco PD. 2018. Reliability in evaluator-based tests: using simulation-constructed models to determine contextually relevant agreement thresholds. *BMC Medical Research Methodology* **18(1)**:1–12 DOI 10.1186/s12874-018-0606-7.

Bermingham A, Smeaton AF. 2009. A study of inter-annotator agreement for opinion retrieval. In: *SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 784–785.

Besacier L, Barnard E, Karpov A, Schultz T. 2014. Automatic speech recognition for under-resourced languages: a survey. *Speech Communication* **56**:85–100 DOI 10.1016/j.specom.2013.07.008.

Bird S, Klein E, Loper E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. Newton: O'Reilly Media, Inc.

Blinov P, Kotelnikov E, Pestov O. 2013. Research of lexical approach and machine learning methods for sentiment analysis. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2013*. Vol. 2. Russian State University for the Humanities, 51–61.

Borodkina O, Sibirev V. 2019. Migration issues in Russian Twitter: attitudes to migrants, social problems and online resources. In: El Yacoubi S, Bagnoli F, Pacini G, eds. *Internet Science*. Cham: Springer International Publishing, 32–46.

Buntoro GA, Adji TB, Purnamasari AE. 2016. Sentiment analysis candidates of Indonesian Presiden 2014 with five class attribute. *International Journal of Computer Applications* **136(2)**:23–29 DOI 10.5120/ijca2016908288.

Chetviorkin I, Braslavskiy P, Loukachevich N. 2013. Sentiment analysis track at ROMIP 2011. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2012*. Vol. 2. Russian State University for the Humanities, 1–14.

Chetvirokin I, Loukachevitch N. 2013. Sentiment analysis track at ROMIP 2012. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2013*. 2.

Chizhik A. 2016. Factors for forming social mood on the basis of the analysis of the emotional coloring of posts in the Russian-language Twitter. In: *New Information Technologies in Automated Systems*. Vol. 19. HSE Tikhonov Moscow Institute of Electronics and Mathematics, 61–64.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Long and Short Papers*. Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Dzogang F, Goulding J, Lightman S, Cristianini N. 2017a. Seasonal variation in collective mood via Twitter content and medical purchases. In: *International Symposium on Intelligent Data Analysis*. Springer, 63–74.

Dzogang F, Lightman S, Cristianini N. 2017b. Circadian mood variations in Twitter content. *Brain and Neuroscience Advances* **1**:2398212817744501 DOI 10.1177/2398212817744501.

Fiok K, Karwowski W, Gutierrez E, Wilamowski M. 2021. Analysis of sentiment in tweets addressed to a single domain-specific Twitter account: comparison of model performance and explainability of predictions. *Expert Systems with Applications* **186**:115771 DOI 10.1016/j.eswa.2021.115771.

Golubev A, Loukachevitch N. 2020. Improving results on Russian sentiment datasets. In: Filchenkov A, Kauttonen J, Pivovarova L, eds. *Artificial Intelligence and Natural Language*. Cham: Springer International Publishing, 109–121.

Hillaire GE. 2021. *Understanding emotions in online learning: using emotional design and emotional measurement to unpack complex emotions during collaborative learning*. Milton Keynes, UK: Open University.

Hillaire G, Rienties B, Fenton-O'Creevy M, Zdrahal Z, Tempelaar D. 2021. Incorporating student opinion into opinion mining. In: *Open World Learning: Research, Innovation and the Challenges of High-Quality Education*. London: Routledge, 171–185 DOI 10.4324/9781003177098-15.

Jamadi Khiabani P, Basiri ME, Rastegari H. 2020. An improved evidence-based aggregation method for sentiment analysis. *Journal of Information Science* **46(3)**:340–360 DOI 10.1177/0165551519837187.

Kanev AI, Savchenko GA, Grishin IA, Vasiliev DA, Duma EM. 2022. Sentiment analysis of multilingual texts using machine learning methods. In: *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. Piscataway: IEEE, 326–331.

Kausar MA, Soosaimanickam A, Nasar M. 2021. Public sentiment analysis on Twitter data during COVID-19 outbreak. *International Journal of Advanced Computer Science and Applications* **12(2)**:415–422 DOI 10.14569/issn.2156-5570.

Kazyulina M, Babii A, Malafeev A. 2020. Emotion classification in Russian: feature engineering and analysis. In: *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 135–148.

Kirilenko AP, Stepchenkova SO. 2017. Sochi 2014 Olympics on Twitter: perspectives of hosts and guests. *Tourism Management* **63**:54–65 DOI 10.1016/j.tourman.2017.06.007.

Koltsova OY, Alexeeva S, Kolcov S. 2016. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2016*. Russian State University for the Humanities, 277–287.

Konstantinov A, Moshkin V, Yarushkina N. 2021. Approach to the use of language models BERT and Word2Vec in sentiment analysis of social network texts. In: *Recent Research in Control Engineering and Decision Making*. Cham: Springer International Publishing, 462–473.

Korablinov V, Braslavski P. 2020. RUBQ: a Russian dataset for question answering over wikidata. In: *International Semantic Web Conference*. Cham: Springer, 97–110.

Kostenetskiy P, Chulkevich R, Kozyrev V. 2021. Hpc resources of the higher school of economics. *Journal of Physics: Conference Series* **1740**:12050 IOP Publishing DOI 10.1088/1742-6596/1740/1/012050.

Kotelnikov E. 2021. Current landscape of the Russian sentiment corpora. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2021*. Russian State University for the Humanities, 433–444.

Kotelnikova A. 2020. Comparison of deep learning and rule-based method for the sentiment analysis task. In: *2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. Piscataway: IEEE, 1–6.

Krippendorff K. 1980. *Content analysis: an introduction to its methodology*. Thousand Oaks: SAGE.

Krippendorff K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research* **30(3)**:411–433 DOI 10.1111/j.1468-2958.2004.tb00738.x.

Kumar P, Raman B. 2022. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks* **150**:392–407 DOI 10.1016/j.neunet.2022.03.017.

Kuratov Y, Arkhipov M. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2019*. Vol. 18. Russian State University for the Humanities, 333–340.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33(1)**:159–174 DOI 10.2307/2529310.

Larsen ME, Boonstra TW, Batterham PJ, O'Dea B, Paris C, Christensen H. 2015. We Feel: mapping emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics* **19(4)**:1246–1252 DOI 10.1109/JBHI.2015.2403839.

Leetaru K. 2019. Is Twitter's spritzer stream really a nearly perfect 1% sample of its firehose? *Forbes* .

Li M, Ch'ng E, Chong AYL, See S. 2018. Multi-class Twitter sentiment classification with emojis. *Industrial Management & Data Systems* **118(9)**:1804–1820 DOI 10.1108/IMDS-12-2017-0582.

Lopatin V, Ulukhanov I. 2017. Russian language. *Languages of the World*. 276–540.

Loukachevitch N, Blinov P, Kotelnikov E, Rubtsova Y, Ivanov V, Tutubalina E. 2015. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2015*. Vol. 2. Russian State University for the Humanities, 3–13.

Lukashevich N, Rubtsova YV. 2016. Sentirueval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2016*. Russian State University for the Humanities, 416–426.

Mozeticar M, Smailović J. 2016. Multilingual Twitter sentiment classification: the role of human annotators. *PLOS ONE* **11(5)**:e0155036 DOI 10.1371/journal.pone.0155036.

Muhammad SH, Adelani DI, Ahmad IS, Abdulmumin I, Bello BS, Choudhury M, Emezue CC, Aremu A, Abdul S, Brazdil P. 2022. NaijaSenti: a Nigerian Twitter sentiment corpus for multilingual sentiment analysis. *Available at https://arxiv.org/abs/2201.08277*.

Pak A, Paroubek P. 2012. Language independent approach to sentiment analysis (LIMSI participation in ROMIP'11). In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2012*. Vol. 11. 37–50.

Pavliy B, Lewis J. 2016. The performance of Twitter's language detection algorithm and Google's Compact Language Detector on language detection in Ukrainian and Russian tweets. *Bulletin of Toyama University of International Studies* **8**:99–106.

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, AL-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, Hoste V, Apidianaki M, Tannier X, Loukachevitch N, Kotelnikov E, Bel N, Jiménez-Zafra SM, Eryigit G. 2016. SemEval-2016 task 5: aspect based sentiment analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, 19–30.

Prata DN, Soares KP, Silva MA, Trevisan DQ, Letouze P. 2016. Social data analysis of Brazilian's mood from Twitter. *International Journal of Social Science and Humanity* **6(3)**:179–183 DOI 10.7763/IJSSH.2016.V6.640.

Pronoza E, Panicheva P, Koltsova O, Rosso P. 2021. Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management* **58(6)**:102674 DOI 10.1016/j.ipm.2021.102674.

Read J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: *Proceedings of the ACL Student Research Workshop*. Ann Arbor, Michigan: Association for Computational Linguistics, 43–48.

Rodina J, Kutuzov A. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, 1037–1047.

Rogers A, Romanov A, Rumshisky A, Volkova S, Gronas M, Gribov A. 2018. RuSentiment: an enriched sentiment analysis dataset for social media in Russian. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 755–763.

Rosstat. 2022. How many people live in Russia: rosstat announced the first results of the census. *Available at https://www.strana2020.ru/novosti/skolko-lyudey-zhivet-v-rossii-rosstat-ozvuchil-pervye-itogi-perepisi/*.

Rubtsova Y. 2013. A method for development and analysis of short text corpus for the review classification task. In: *Trudy XV Vserossiiskoy Naychnoy Konferencii RCDL'2013*. 269–275.

Salminen JO, Al-Merekhi HA, Dey P, Jansen BJ. 2018. Inter-rater agreement for social computing studies. In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 80–87.

Smetanin S. 2017. The program for public mood monitoring through Twitter content in Russia. *Proceedings of the Institute for System Programming of the RAS* **29(4)**:315–324 DOI 10.15514/ISPRAS-2017-29(4)-22.

Smetanin S. 2020a. The applications of sentiment analysis for Russian language texts: current challenges and future perspectives. *IEEE Access* **8**:110693–110719 DOI 10.1109/ACCESS.2020.3002215.

Smetanin S. 2020b. Toxic comments detection in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2020*. Russian State University for the Humanities.

Smetanin S, Komarov M. 2019. Sentiment analysis of product reviews in Russian using convolutional neural networks. In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. Vol. 1. IEEE, 482–486.

Smetanin S, Komarov M. 2021a. Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management* **58(3)**:102484 DOI 10.1016/j.ipm.2020.102484.

Smetanin S, Komarov M. 2021b. Share of toxic comments among different topics: the case of Russian social networks. In: *2021 IEEE 23rd Conference on Business Informatics (CBI)*. Vol. 2. Piscataway: IEEE, 65–70.

Szczepański M, Pawlicki M, Kozik R, Choraś M. 2021. New explainability method for BERT-based model in fake news detection. *Scientific Reports* **11(1)**:1–13 DOI 10.1038/s41598-021-03100-6.

Szmigiera M. 2022. The most spoken languages worldwide in 2022. *Available at https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/*.

Ten Thij M, Bhulai S, Kampstra P. 2014. Circadian patterns in Twitter. In: *IARIA Data Analytics 2014*. 12–17.

Tripto NI, Ali ME. 2018. Detecting multilabel sentiment and emotions from Bangla YouTube comments. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. Piscataway: IEEE, 1–6.

Wang Y, Callan J, Zheng B. 2015. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Transactions on the Web* **9(3)**:1–23 DOI 10.1145/2746366.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. 2020. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 38–45.

Zueva N, Kabirova M, Kalaidin P. 2020. Reducing unintended identity bias in Russian hate speech detection. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 65–69.

PeerJ Computer Science

Is gender-based violence a confluence of culture? Empirical evidence from social media

Rimjhim and Sourav Dandapat

Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India

ABSTRACT

Gender-based violence (GBV) has been plaguing our society for long back. The severity of GBV has spurred research around understanding the causes and factors leading to GBV. Understanding factors and causes leading to GBV is helpful in planning and executing efficient policies to curb GBV. Past researches have claimed a country's culture to be one of the driving reasons behind GBV. The culture of a country consists of cultural norms, societal rules, gender-based stereotypes, and social taboos which provoke GBV. These claims are supported by theoretical or small-scale survey-based research that suffers from under-representation and biases. With the advent of social media and, more importantly, location-tagged social media, huge ethnographic data are available, creating a platform for many sociological research. In this article, we also utilize huge social media data to verify the claim of confluence between GBV and the culture of a country. We first curate GBV content from different countries by collecting a large amount of data from Twitter. In order to explore the relationship between a country's culture and GBV content, we performed correlation analyses between a country's culture and its GBV content. The correlation results are further re-validated using graph-based methods. Through the findings of this research, we observed that countries with similar cultures also show similarity in GBV content, thus reconfirming the relationship between GBV and culture.

Subjects Human-Computer Interaction, Data Science, Social Computing, Spatial and Geographic Information Systems, World Wide Web and Web Science

Keywords Computational social science, Social computing, Ethnography, Gender based violence, Twitter, Correlation, Hofstede Cultural Dimension, Culture, Sexual violence, Physical violence

Submitted 15 February 2022
Accepted 30 June 2022
Published 29 July 2022

Corresponding author
Rimjhim, rimjhim.pcs16@iitp.ac.in

Academic editor
Nageswara Rao Moparthi

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.1051**

Copyright
2022 Rimjhim and Dandapat

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Gender-based violence (GBV) is one of the most heinous and age-old violations of human rights (https://www.who.int/news-room/fact-sheets/detail/violence-against-women). GBV is evident across all parts of the globe (https://www.undp.org/content/undp/en/home/blog/2018/violence-against-women-cause-consequence-inequality.html), and it has been plaguing our society for a long back. The condition is so severe that one in three women is reported to have faced GBV (https://www.who.int/news/item/09-03-2021-devastatingly-pervasive-1-in-3-women-globally-experience-violence). With alarming instances of GBV around the world, social and governmental organisations are taking rigorous preventive measures. The quest to deliver effective preventive measures has triggered research to understand the causes and factors of GBV to provide effective

preventive measures. Research in this field have found that cultural norms which comprise of societal stigma, gender-based rules, and societal prejudices are major factors that contribute to GBV (*Jewkes, Jama-Shai & Sikweyiya, 2017*; *Elischberger et al., 2018*; *Raj & Silverman, 2002*; *Jewkes, Levin & Penn-Kekana, 2002*; *Bishwajit, Sarker & Yaya, 2016*). GBV is pervasive across all social, economic, and national strata (*Dartnall & Jewkes, 2013*), but the type of GBV, the intensity of GBV, people's reactions, and opinions for any GBV event is not the same across the globe. For example, acid attacks are a form of revenge in developing countries arising because of refusal of a marriage proposal or a love proposal, or land disputes (*Bahl & Syed, 2003*). However, in South America, the same acid attack results from poor relationships and domestic intolerance toward women (*Guerrero, 2013*). The context of GBV changes with the country, and this change is known to be an outcome of persisting culture in a country (*Abrahams et al., 2014*; *Fulu & Miedema, 2015*; *Alesina, Brioschi & La Ferrara, 2016*; *Perrin et al., 2019*; *Stubbs-Richardson, Rader & Cosby, 2018*). The World Health Organization (WHO) has studied cultural norms of many countries leading to various forms of GBV (https://apps.who.int/iris/bitstream/handle/10665/77936/9789241500845_eng.pdf?sequence=1&isAllowed=y). The global organization World Bank also pronounced to work on such cultural and social norms to curb GBV (https://thedocs.worldbank.org/en/doc/656271571686555789-0090022019/original/ShiftingCulturalNormstoAddressGBV.pdf).

However, these researches claiming cultural norms as a driving factor behind GBV are based on cognitive studies which require significant intervention from social and cultural experts. The claims presented in these works are based upon long-term manual discerning of GBV events occurring in countries of different cultures. These researches are dependent upon survey/questionnaire-based data which can be collected only in a limited amount and can also suffer from several biases. Thus, past research lacks a large-scale, data-driven empirical research to verify the confluences between culture and GBV.

In this article, we take a step to answer the research question "Is gender-based violence a confluence of culture?" by experimenting with large-scale social network data. The use of social network data for research around GBV is a non-conventional way to dive into the finer details of GBV. Our research analyses GBV from the lens of culture. This research is useful for social workers, policy-makers, governments, and other organizations working for the welfare of women and society (*Kim, 2021*). Additionally, the findings of this research can help in planning more efficient and targeted GBV policies and awareness campaigns. Social network data has already become a substitute for survey data for numerous applications. Recently, social network data has also gained much utility for research related to GBV (*Hansson, Sveningsson & Ganetz, 2021*; *Liu et al., 2019*; *Chowdhury et al., 2019*; *Hassan et al., 2020*). Online content contains a rich spectrum of information pertaining to user opinions/reactions, ongoing news/events (*Blake et al., 2021*), and many more (*Nikolov et al., 2015*; *Pal et al., 2018*). Thus, online content is not only a mere content but a real-time proxy for user behaviour. For this research, we consider online content related to GBV from different countries as a representative of user reactions and perspectives towards GBV. We design experiments to check for the content similarity between countries with similar cultures. Towards this goal, we perform the correlation analysis between

content distance and cultural distance between countries. Further, to validate results from the correlation analysis, we also performed graph analyses. In graph analyses, we create graphs with countries as nodes and different types of distances (content distance and cultural distance) between countries is used for building edges. These graphs are compared using various graph comparison metrics.

On experimentation with Twitter content from different countries, we find a statistically significant positive correlation between GBV content distance and cultural distance. We also observed a higher similarity between the GBV content graph and the culture graph. Thus, through the findings of this research, we observe that the countries which are similar in culture also show higher similarity in GBV content. This observation is consistent with correlation analyses and graph analyses. From this observation, we can conclude that there are traces of culture in GBV content which justifies the claim of confluences of culture on GBV. The contributions of the current research can be summarized as follows:

- In this research, we explore evidence of confluence between GBV and the culture by means of an empirical study conducted over a large dataset created naturally over a long period of time on social media.
- The results obtained from this research justify the hypothesis that GBV is a confluence of culture. This hypothesis has not been tested in past literature using uncensored and unbiased social media data.
- All the experiments conducted in this research are extended to different categories of GBV and generic online content. Further, all the six dimensions of culture are also investigated. Thus, we provide a holistic analysis.
- The findings in this research are supported by correlation analyses as well as graph-based analyses. Thus, making our claims more robust.

The rest of the article is organized as follows. 'Related Works' details relevant past literature related to this research. 'Dataset and Processing' describes the collected data. 'Methodology' elaborates on the methodology of our experiments, and 'Results' shows all the results and analyses. 'Discussions' discusses the implications and limitations of the work. Finally, 'Conclusion and Future Work' concludes our work with possible future works.

RELATED WORKS

This research is based upon three broad areas of related works i. The relation between GBV and culture ii. Social Media Content as a source of Data and iii. GBV through social media.

GBV and culture

GBV is a social ill evident across all the countries irrespective of their economy, language, and demography. However, with country, the type of GBV, its intensity, and the reaction of people vary (*Fakunmoju, Bammeke et al., 2017*). For example, in the USA, dating violence is more common than in Africa where there are comparatively lesser instances of dating violence (*Johnson et al., 2015*). On the other hand, in Africa, intimate partner violence is more prominent as compared to North America (https:

//apps.who.int/iris/bitstream/handle/10665/85239/9789241564625_eng.pdf). This implies that the same GBV is represented differently in a different country. This implies that GBV is a global evil but the context of GBV changes with the country. There have been many research to understand the causes and factors leading to GBV (*Jewkes, Levin & Penn-Kekana, 2002*; *Jewkes, Jama-Shai & Sikweyiya, 2017*; *Marine & Lewis, 2020*). These works have claimed that a country's culture can characterize the persisting GBV in the country. Every culture has norms, prejudices, and societal rules that design the behaviour of people towards GBV. For example, in Malawi, the concept of polygamy and dowry is evident in the culture, and these perpetuate GBV in Malawi (*Bisika, 2008*). Similar research in many other countries like UK (*Aldridge, 2021*), Ethiopia (*Le Mat et al., 2019*), Cambodia (*Palmer & Williams, 2017*), and many other countries (*Djamba & Kimuna, 2015*; *Raj & Silverman, 2002*) have highlighted cultural norms which lead to one or other form of GBV. Not only in research but global organisations like WHO (https://www.who.int/violence_injury_prevention/violence/norms.pdf), World Bank (https://thedocs.worldbank.org/en/doc/656271571686555789-0090022019/original/ShiftingCulturalNormstoAddressGBV.pdf) have also highlighted the cultural norms of many countries that influence GBV. The socio-cultural impact is so intense that people even justify instances of GBV as a form of the social norm which cannot be questioned (*Piedalue et al., 2020*). However, these claims are supported by mere examples and small-scale interview-based data. Thus, the research community lags a data-driven research that justifies the claim with sufficient empirical results. In this research, we do a large-scale analysis of social network content to find evidence of confluence between the culture of a country and GBV. Next, we present the role of social media content in bridging the gap of data for various research.

Social media content

Social media has now become the new language of people, and this has generated a massive amount of data for various research. Social media data has removed the bottleneck of data requirements in numerous applications such as urban computing (*Silva et al., 2019*), cultural computing (*Wang et al., 2017*), personality computing (*Samani et al., 2018*) and many more. Social media content also plays a huge role in understanding people's views and sentiments during the COVID-19 pandemic (*Malagoli et al., 2021*). Social media content has already substituted the tedious, time consuming, biased and under-represented survey-based data and has unlocked possibilities for research in many other directions. The utility of social media content increases with the availability of location-tagged data. The location-tagged online content has been used in numerous ethnographic research (*Abdullah et al., 2015*), cultural research (*Cheke et al., 2020*), and sociological research (*Stewart et al., 2017*) in recent days. Social media content is not mere data, but it captures several dimensions of human interests, psychology, and behavior. There have been works which found that the social media content reflects the real-world properties as well. *García-Gavilanes, Mejova & Quercia (2014)*; *Garcia-Gavilanes, Quercia & Jaimes (2013)* have found that interaction and usage of social networks are dependent on social, economic, and cultural aspects of users. Thus, the real-world behavior of people is also mirrored in social networks. This

utility of social media content motivates us to examine the relationship between GBV and a country's culture through analysis of social media data. For this research, we use Twitter data related to GBV from different countries.

Measuring culture

Culture is an amalgamation of thoughts, beliefs, potential acts, and a lot more. A number of definitions of culture are available from previous works (*Hofstede, Hofstede & Minkov, 2005*). One of many definitions of culture is *"a fuzzy set of assumptions and values, orientations to life, beliefs, policies, procedures and behavioural conventions that are shared by a group of people"* (*Spencer-Oatey & Franklin, 2012*). Culture plays a vital role in many spheres of life, such as behaviour (*Huang, Beshai & Yu, 2016*) economy (*Herrero, Jiménez & Alcalde, 2021*), language (*Sazzed, 2021*), attitude (*Shin, Chotiyaputta & Zaid, 2022*). In order to ease out research based on culture, there have been several quantification of culture. Hofstede has done one such comprehensive quantification. Hofstede (*Hofstede, Hofstede & Minkov, 2005*) defines culture in terms of six parameters (*Power Distance, Uncertainty Avoidance, Individualism, Masculinity, Long Term Orientation, Indulgence*) and quantifies each one for different countries of the world. An extensive set of previous works use Hofstede Dimensions to quantify culture (*García-Gavilanes, Mejova & Quercia, 2014*; *Garcia-Gavilanes, Quercia & Jaimes, 2013*; *Prakash & Majumdar, 2021*). For this research, we also used Hofstede's dimensions which have been used in a huge number of research to measure culture.

GBV through the lens of social media

Social media provides an uncensored and user-friendly medium for expressing views and opinions (*Puente, Maceiras & Romero, 2021*). With this, social media has become a platform for self-expression as well as for conducting online campaigns (*Martínez, Pacheco & Galicia, 2021*). There have been many campaigns on social media related to GBV like the #metoo, #Notokay, #StateOfWomen, #HeForShe and many more (*Karuna et al., 2016*). These campaigns and freedom of expression on social media have generated huge data related to GBV. The recent campaign of #Metoo observed an unprecedented response from all around the globe, thus, generating huge data related to GBV. And the event was followed by a sudden upsurge in research related to GBV using the generated data.

Thus, the data availability of social media has helped in many recent works related to GBV, which have delivered a multitude of interesting findings (*Moitra, Ahmed & Chandra, 2021*; *Razi, 2020*; *Pandey et al., 2018*; *Khatua, Cambria & Khatua, 2018*). Moreover, location-tagged social media data also assist in several cross-cultural studies related to GBV (*Purohit et al., 2015*; *Starkey et al., 2019*). In this article, we also used social media Twitter data from different countries of the world as a source of data.

METHODOLOGY

In this section, we first give details of the collected dataset and its processing. Further, in the section, we elaborate on the methodology used to understand the relation between GBV and culture. We process the country-level tweets from different categories for correlation

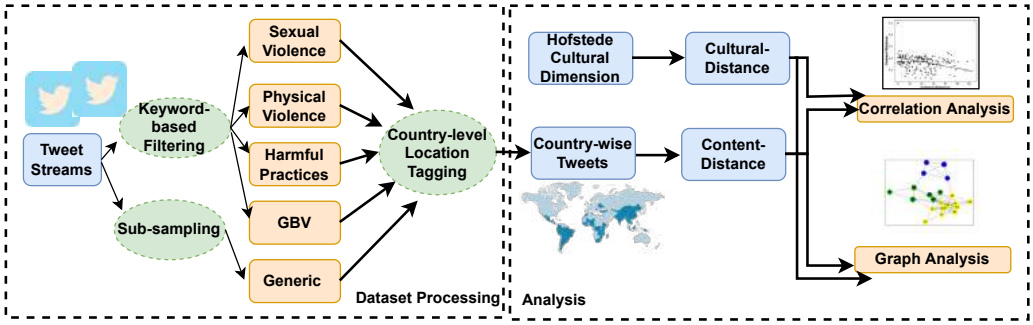


Figure 1 **An overview of our proposed methodology.**
Full-size DOI: 10.7717/peerjcs.1051/fig-1

analysis and graph analysis w.r.t culture of different countries. The flow of methodology is represented in Fig. 1. Next, we explain the details of data collection and its processing to obtain country-wise GBV tweets.

Dataset and processing

We use public streams of Twitter data collected using the Twitter Streaming API. We procured 1% of public tweets provided by the API for a period of two years and five months (1st July 2016–25th Nov 2018). We remove all the duplicate tweets and retweets from the collected data as these do not add any new information (*Cheke et al., 2020*). From the collected tweets, we extracted GBV related tweets using a keyword matching approach as described next.

GBV Tweet extraction

UNFPA (United Nations Population Fund) domain experts have proposed three categories of GBV, namely *sexual violence*, *physical violence*, and *harmful practices*. They have also provided unique keywords related to each category of GBV, which have been used frequently in past literature for GBV related research (*Purohit et al., 2015*; *ElSherief, Belding & Nguyen, 2017*). Table 1 shows a total of 81 keywords constituting 29, 25, and 27 keywords from sexual violence, physical violence, and harmful practices respectively. We use the same keywords to extract relevant tweets from all three categories. The keyword set provided by UNFPA is very precise and can contain multi-words. Our methodology for extracting tweets for a particular keyword is based on the presence of the keyword in a tweet. If all the words of a multi-word keyword are present in a tweet regardless of order, we consider it a match. For example, for the category sexual violence, *sexual assault* is a related keyword with two words. If a tweet contains both the words *sexual* and *assault*, we consider it a match. For the cases where a tweet matches more than one category, we consider the tweet in both categories of GBV. This approach has been used in previous works in order to deliver high-precision data (*Purohit et al., 2015*). From the keywords related to each category of GBV, we extract tweets and create a tweet dataset from three categories, namely the sexual violence dataset, physical violence dataset, and harmful practices dataset, with a total of 0.83*million*, 0.53*million*, and 0.66*million* tweets, respectively. Further, we combined

Table 1 Set of keywords to identify tweets from different categories.

Category	Relevant keywords
Sexual violence	sexual assault, sexual violence, woman/women/girl/female harass, woman/women/girl/female attacked, boyfriend/boyfriend assault, stalking woman/women/girl/female, groping woman/women/girl/female, sexual/rape victim, gang rape, victim blame, sex predator, woman/women/girl/female forced
Physical violence	woman/women/girl/female beat up, woman/women/girl/female acid attack, woman/women/girl/female violence, woman/women/girl/female punched, woman/women/girl/female attacked, gender/domestic violence, intimate partner violence, physical abuse/violence
Harmful practices	child/children/underage/forced marriage, sex/child/children trafficking, woman/women/girl/female trafficking, child molestation/bride/sex, child violence/abuse/bullying/beat, spouse abuse, sex/women/forced slave, female genital mutilation (fgm), early marriage, pedophilia, human trafficking, woman abuse
GBV	All the keywords from sexual violence, physical violence, and harmful practices

Table 2 Descriptive statistics of collected data.

Data category	Before location tagging	After location tagging
Sexual violence	836,497	681,537
Physical violence	534,707	433,560
Harmful practices	659,666	522,844
GBV tweets	2,030,874	1,637,941
Generic tweets	42,445,234	36,689,133

all three category tweets to create a GBV tweet dataset containing more than two million tweets.

Generic Tweet dataset

We created another dataset, the *generic tweet dataset*, to provide a better context of comparison with other categories of the dataset. This dataset is used for drawing inferences from GBV categories dataset w.r.t a generic dataset. For creating this dataset, we borrowed the methodology of *ElSherief, Belding & Nguyen (2017)*. Our collected data is for a very long period, resulting in around 4 billion tweets. We extracted a random 1% sub-sample of total collected tweets as a *generic tweet dataset*. To eliminate duplicate content here as well, we removed tweets which are duplicates and retweets. Details of generic tweets data are given in Table 2.

Country-level location tagging

There are many indicators of location in a tweet, such as *geotags*, *time zone*, and *profile location*. Adopting the location indicators of *Annamoradnejad et al. (2019)* for tagging

[1]This is a measure of the level of acceptance of unequal power in society.

[2]This is a measure of rights and concerns of each person rather than for a group or community.

[3]This is a measure of the distribution of gender-based roles in society.

[4]This is a measure of likeliness that people avoid uncertainty.

[5]This parameter measures the characteristics of perseverance and futuristic mindset among people.

[6]This measures the degree of fun and enjoyment a society allows.

each tweet to a location, we use a three-level hierarchy of location indicative according to their accuracy levels (*Kulshrestha et al., 2012*). The first one is geotag, which gives the most accurate location information. If a geotag is available, then we use it for location tagging, and if it is not present, we look for the time zone data. Time zone is also an accurate way to tag country-level locations. A time zone data directly contains the user's time zone in the form of the corresponding country name. For the cases where even time zone information is not available, then we look for the next location information in the hierarchy, *i.e.,* location field mentioned in the user profile. Geotags and time zone contain exact country names, which can be directly mapped to a country. User profile location is an unstructured text location field that requires further processing to get country information. For this, we use the approach used by *Dredze, Osborne & Kambadur (2016)* where city names present in the user profile location are mapped to corresponding country names based on the Geoname (https://pypi.org/project/geonames/) world gazetteer. We borrow the list of required countries from *Annamoradnejad et al. (2019)* where authors have used a list of 22 countries, namely *Arab countries, Argentina, Australia, Brazil, Canada, China, Colombia, France, Germany, India, Indonesia, Iran, Italy, Japan, Korea, Philippines, Russia, Spain, Thailand, Turkey, UK (United Kingdom), USA(United States of America)*. *Arab Countries* is a group of countries with a similar culture, so we merged tweets from all *Arab Countries*. There were very few tweets from Korea (170), so we discarded Korea from the list of considered countries and limited our research to the remaining 21 countries, each having more than 3,000 tweets. We apply the same location tagging scheme to all the GBV tweets and generic tweets. The complete data statistics are shown in Table 2 for all the categories of tweets.

We present the evaluation methodology and evaluation results of GBV tweets extraction and location tagging in 'Results'. In this research, we want to explore the relationship between GBV online content and culture of a country. For this, we perform two analyses i. Correlation Analyses and ii. Graph Analyses. In correlation analysis, we correlate culture and its dimensions with different categories of online content in order to understand their relationship. In graph analyses, we create country graphs on the basis of parameters correlated in correlation analyses like content, and culture, which are compared using multiple graph comparison metrics in order to re-assure the observed relationships from correlation analyses. Next, we discuss the methodology used for these analyses.

Correlation analysis

In order to find a relation between the culture of a country and GBV, we calculate cultural distance and content distance between each pair of countries as detailed next.

Cultural distance

We quantify the cultural distance between two countries using cultural dimensions proposed by Geert Hofstede (*Hofstede, Hofstede & Minkov, 2005*). Geert Hofstede administered a huge survey among people from different countries to measure the difference in the way they behave. He has quantified six dimensions of culture (*power distance*[1], *individualism*[2], *masculinity*[3], *uncertainty avoidance*[4], *long-term orientation*[5],

indulgence[6]) for different countries in values ranging between 0–120. In order to measure cultural distance between two countries, we adopt the formulation of (*Annamoradnejad et al., 2019*) where authors use the euclidean distance between two countries to measure the cultural distance. The cultural distance can be formulated as shown in Eq. (1), where $|D|$ is the total number of dimensions, $d^i_{c_1}$ and $d^i_{c_2}$ are the values of dimension d^i for countries c_1, c_2 respectively.

$$Cultural\ Distance(C_1, C_2) = \sqrt{\sum_{i=1}^{|D|}(d^i_{c_1} - d^i_{c_2})^2} \tag{1}$$

We also calculate the distance between countries on the basis of each dimension of culture proposed by Hofstede. For example, *power distance* is one of the dimensions of culture, and we need to calculate the distance between two countries according to *power distance*. For this also, we use euclidean distance, but since there is only one parameter, this becomes equivalent to $|d_{c_1} - d_{c_2}|$. For further analyses, we calculate the cultural distance for each pair of countries on the basis of culture and six dimensions of culture.

Content distance

Online content related to a particular topic from a particular country captures country-level user comments and discussions on that topic (*Cheke et al., 2020*). In order to measure the difference between contents from two countries, we measure the *content distance* between two countries using Jaccard Similarity. First, all the tweets from each country are pre-processed to generate country-wise tweet tokens, details of which are given next.

[7] Jaccard similarity is a popular metric for measuring content similarity. We also experimented with Cosine similarity, another popular metric and found similar result patterns.

Tweet preprocessing

We adopt the pre-processing settings of *Cheke et al. (2020)* to generate tweet tokens from each country. We first remove URLs, mentions, punctuation, extra spaces, stop words, and emoticons. Online acronyms and short forms are expanded using NetLingo (https://www.netlingo.com/acronyms.php). For hashtags, we removed the symbol # and kept the remaining word. Spelling and typos are corrected using Textblob (https://textblob.readthedocs.io/en/dev/). We also transliterated non-English words to English to reduce inconsistencies in language. Lastly, we tokenized each tweet using NLTK (Natural Language Toolkit). Extracted tokens from all the tweets of a country are merged to create country-wise tweet tokens. Next, for each pair of countries, we calculate *content distance* using the formula shown in Eq. (2), where C_1, C_2 are the set of all the tweet tokens of countries C_1 and C_2, respectively, and $\frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$ is the Jaccard Similarity.[7]

$$ContentDistance(C_1, C_2) = 1 - \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}. \tag{2}$$

We have a total of 5 categories of online content *i.e.,* sexual violence, physical violence, harmful practices, GBV, and generic content. We apply the same methodology to extract country-wise tweet tokens from each category of online content.

Correlation

A correlation helps in understanding the relationship between two variables. *Pearson correlation* and *Spearman correlation* are two popular metrics for correlation. To establish

robustness in our findings, we use both *Pearson correlation*, and *Spearman correlation* for calculating the association between *content distance* and *cultural distance*. *Pearson correlation* captures the linear relationship between two variables and *Spearman correlation* captures the monotonic relationship between two variables. Both the correlation metrics give correlation values in the range of $(-1$ to $+1)$. A positive correlation value indicates that content similarity is higher for countries having higher cultural similarity and a negative correlation indicates vice versa. For calculating the correlation, we calculated content distance and corresponding cultural distance for each country pair (a total of ${}^{n}C_{2}$ pairs, if there are n countries). For exhaustive correlation analysis, we measured multiple correlations by keeping one correlation variable as different types of content (sexual, physical, harmful, GBV, and generic tweets) and another variable as six dimensions of culture.

P-value

For measuring the fitness of a correlation, we calculated the p-values for each correlation using the python library SciPy (https://docs.scipy.org/doc/scipy-0.14.0/reference/stats.html). A p-value represents the measure of occurrence of the correlation between two data samples by chance.

Graph analysis

Our objective is to compare GBV related content to a country's culture. To this end, we created country graphs where edge weights are decided on the basis of different distances in terms of GBV content and culture, as mentioned in 'Correlation Analysis'. For detailed analyses, we created multiple weighted graphs among countries with a different edge parameters. Finally, we compared created graphs using multiple graph distance metrics and graph clustering.

Country graph: A country graph created in this research is an un-directed, weighted graph $G = (C, E)$, where C denotes the nodes of the graph, which are countries, and E denotes the set of edges between countries. For all the graphs in this research, the set of nodes C and the set of edges E are the same. The only difference is in the weights of the edges. Next, we describe the creation of edges in the required graphs.

GBV content graph: This graph captures the relationship between countries according to GBV content distance. In GBV content graph $G_{gbv} = (C, E)$, the weights of the set of edges E are decided on the basis of the *content distance* score between two countries on the basis of GBV tweets. Here, GBV tweets are used for calculating content distance. We also create *sexual violence graph*, *physical violence graph*, and *harmful practices graph* where for assigning edge weights, we calculate content distance on sexual violence tweets, physical violence tweets, and harmful practices tweets, respectively.

Generic content Graph: This graph captures the relationship between countries and generic content. In the generic content graph $G_{rand} = (C, E)$, the weights of the edges are assigned according to the *content distance* score between two countries on the basis of generic tweet data.

Cultural graph: This graph captures the cultural relationship between countries. In the cultural graph $G_{cult} = (C, E)$, the weights of the edges are decided by the value of cultural distance calculated using Eq. (1).

Graph Pre-processing: For all the graphs $G = (C, E)$, there is an edge between any pair of countries with a weight creating a complete graph. Further, all the created graphs have a different range of values of edge weight. For example, for GBV tweets graphs, edge weights will lie in the range (0,1), but for the culture graph, the values of weights can range from (0–120). To ensure consistency, we upscale edge weights in the range of (0,1) to a range of (0–120). Next, we pruned edges that are unimportant, *i.e.,* whose weight is lesser than the median of all the edge weights. Thus, keeping only the important, *i.e.,* higher edge weight edges in the graph. Before pre-processing, each graph is a complete graph with the same edges in all the graphs, but after pre-processing, each graph is a non-complete graph with only important edges resulting in different edges in each graph. The same is applied to all the graphs, and the final pre-processed graph is a weighted, un-directed, and non-complete graph.

We also mention that all the distances (content/culture) used to decide edge weight in all the graphs follow the axioms of distance (*Kosub, 2019*).

Graph comparison metrics

For comparing two graphs, past literature has proposed a number of metrics depending upon the type of graphs (*Tantardini et al., 2019*; *McCabe et al., 2020*). In this research, we have graphs with node correspondence, *i.e.,* same nodes in every graph. Additionally, our graphs are un-directed and weighted. For the purpose of graph comparison, we use multiple graph distance metrics to calculate the distance between two graphs. Graph distance shows how dissimilar the two graphs are. For calculating graph distance, we used the Python library *netrd* (https://netrd.readthedocs.io/en/latest/distance.html). Next, we describe metrics used in our research to calculate graph distance.

Distances:

- Quantum JSD: Quantum Jensen–Shannon Divergence (*De Domenico & Biamonte, 2016*) compares two weighted and undirected graphs by finding the distance between spectral entropy of density matrices.
- Degree divergence: This method (*Hébert-Dufresne, Grochow & Allard, 2016*) compares the degree distribution of two graphs. This methodology is applicable to weighted as well as unweighted graphs but only undirected graphs.
- Jaccard distance: Jaccard distance (*Oggier & Datta, 2021*) is applicable to only unweighted graphs, and its value depends on the number of common edges in the two compared graphs. For applying to our graphs, we coerced weighted graphs into unweighted ones by removing weights from all the graphs.
- Hamming distance: Hamming distance is one of the popular techniques for measuring the distance between two unweighted graphs. This is a measure of element-wise disagreement between the two adjacency matrices of the graphs. We applied Hamming distance to our graphs by coercing weighted graphs to unweighted ones by simply removing the weights.

- HammingIpsenMikhailov: This method is an enhanced version of Hamming Distance which takes into account the disagreement between adjacency matrices and associated laplacian matrices. This is applicable to weighted and undirected graphs.
- Frobenius: This is an adjacency matrix level distance that calculates L2-norms of the adjacency matrices.
- NetLSD: A metric for measuring graph distance based on spectral node signature distributions for unweighted graphs. For this, we coerced our graphs to unweighted ones by removing the weights.

Graph clustering

A graph clustering algorithm clusters similar nodes in different groups. If two graphs are similar, then their clusters will also be similar. In order to compare the GBV graph and the generic graph with the culture graph, we used the Louvain community detection algorithm (*De Meo et al., 2011*). Louvain community detection algorithm is a clustering algorithm for nodes of a weighted graph where nodes are clustered on the basis of modularity between the nodes. Here the number of clusters was decided by the algorithm only.

RESULTS

In this section, we first provide validation results for our proposed methodology of GBV tweet filtering and location tagging. Then we present the results and insights of correlation analyses and graph analyses in order to understand the correspondence between GBV and culture.

GBV tweet extraction and error analysis

GBV tweet extraction is accomplished by tagging tweets using a keyword matching process. Following the keyword match verification methodology of *Cheke et al. (2020)*, we employed three graduate annotators to manually tag the GBV category. Annotators were provided with a sample of tweets without any category information and were asked to manually tag each tweet to one or more categories of GBV (sexual violence, physical violence, harmful practices) with their own understanding and external online resources. Annotators were provided with a basic definition of GBV and its categories. For the purpose of validation, we created a balanced and shuffled sample of 6,000 tweets with 2,000 tweets from each category of GBV. For each tweet annotated by the three annotators, we select the majority category as the final category. Tweets that do not have any majority category are discarded. Considering the category tagged by annotators as the actual categories, we calculate the precision value of our keyword matching methodology for each category of GBV. The precision value for sexual violence is found to be 0.97, for physical violence 0.96, and for harmful practices to be 0.98.

Table 3 shows a few example tweets and tagged GBV categories. Examples 1–9 shows matching keywords and the tagged GBV category of the tweets from all three categories of GBV. Example 10–11 show tweets that contain keywords from more than a category of GBV. These tweets are kept in all the matching categories. In examples 12–13, keyword matching results in the wrong tagging of tweets because of contextual differences in tweets.

Table 3 Example tweets with their matching keywords and tagged GBV category.

S.No.	Example tweets	Category
1	Why is **harassment** an automatic career hazard for a **woman** receiving any amount of professional attention?	Sexual violence
2	**Girls** are **forced** to sleep and authorities are POWERLESS. Europe is dead.	Sexual violence
3	Yesterday I asked my daughter's schools to stop **slut-shaming** and **victim-blaming** girls. It went viral.	Sexual violence
4	#Berlin metro attacker who kicked **woman** down stairs in random act of **violence** detained	Physical violence
5	laws don't cause divorce, **domestic violence** does	Physical violence
6	Some **girls** are **beaten up** by their boyfriends and stick around saying Ï see something in him.	Physical violence
7	Gather round children, I'm doing a thread on how this society sexualizes **underage girls**. Leggo.	Harmful practices
8	If you suspect a **child** is being **abused** You have a moral duty to report it	Harmful practices
9	A 12 year old **child bride** taking photos in her wedding dress. Can you imagine it?	Harmful practices
10	For most of these **women**, a history of **sexual abuse** and **childhood** trauma dragged them into prostitution	Multiple
11	She has written books on **sexual abuse**, **child molestation**, **domestic violence**.	Multiple
12	**Woman** With Too Much Makeup Mistaken As Clown; **Attacked** By Angry Mob	Physical (Error)
13	all I want for my children is happiness! I don't care what their **sexual** preference Is.. No one can **force** them	Sexual (Error)

As we can see in example tweet 12, the keywords *woman* and *attacked* belong to physical violence keywords, and hence the tweet is wrongly classified in the physical violence category. There are only a few such errors in GBV tweet category tagging arising because of changes in the context of tweets.

Evaluation of location tagging

We have a three-level hierarchy(time zone, geotags, profile location) of location tagging. Location tagging from time zone and geotags is completely accurate. For evaluating location tagging from profile location a random sample of 10,000 tweets are given to three independent graduate annotators who were asked to manually tag a country-level location from their own understanding using online gazetteers and searches. The majority country name is selected as the final tagged country name. The profile location field with no majority among annotators is discarded. Considering the country tagged by manual annotation as the actual country, we obtained a precision score of 0.94.

Correlation analysis

Tables 4 and 5 show the results of *pearson correlation* and *spearman correlation* of different types of online content with culture and its parameters. From the tables, we can draw the following observations.

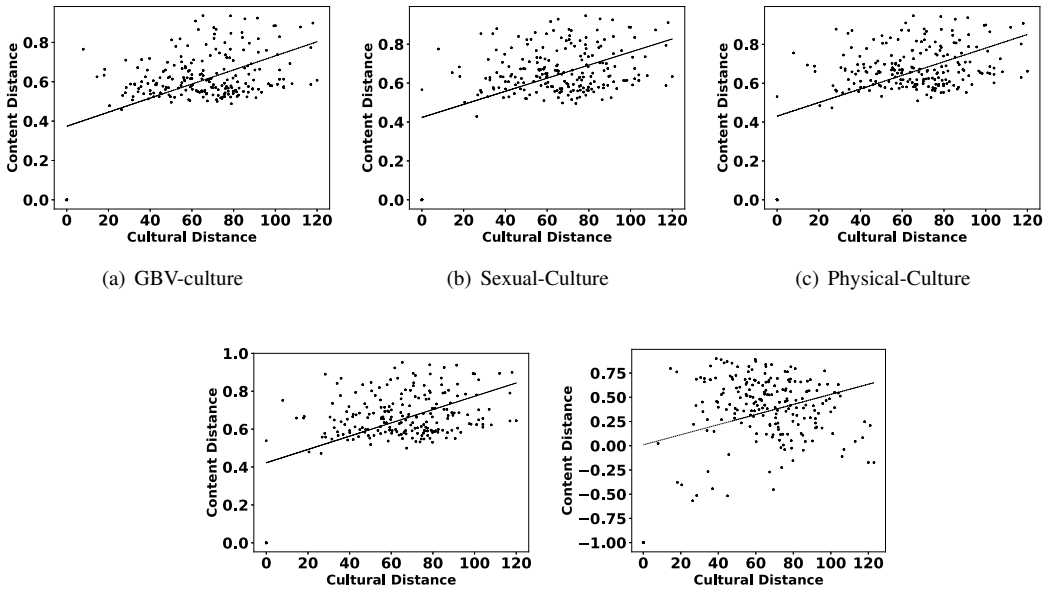
Table 4 Pearson correlation coefficient and *p*-values between different cultural distances and different content distances. All the correlations all calculated on a sample of 210 country pairs, since there are 21 countries. Degree of freedom for the correlation analyses is 208.

	GBV tweets		Sexual violence		Physical violence		Harmful practices		Generic tweets	
	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value
Cultural distance	0.55	0.001	0.51	0.001	0.53	0.001	0.53	0.001	0.33	0.001
Power distance	0.27	0.001	0.30	0.001	0.30	0.001	0.32	0.001	0.12	0.01
Masculinity	0.16	0.001	0.16	0.001	0.18	0.001	0.16	0.001	0.17	0.01
Uncertainty avoidance	0.36	0.001	0.34	0.001	0.31	0.001	0.33	0.001	0.23	0.001
Long-term orientation	0.17	0.001	0.16	0.001	0.19	0.001	0.18	0.001	0.21	0.001
Indulgence	0.34	0.001	0.30	0.001	0.32	0.001	0.30	0.001	0.13	0.01
Individualism	0.45	0.001	0.40	0.001	0.42	0.001	0.42	0.001	0.15	0.001

Table 5 Spearman correlation coefficient and *p*-values between different cultural distances and different content distances. All the correlations all calculated on a sample of 210 country pairs, since there are 21 countries. Degree of freedom for the correlation analyses is 208.

	GBV tweets		Sexual violence		Physical violence		Harmful practices		Generic tweets	
	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value
Cultural distance	0.28	0.001	0.24	0.001	0.25	0.001	0.28	0.001	0.03	0.43
Power distance	0.17	0.001	0.23	0.001	0.21	0.001	0.25	0.001	0.01	0.82
Masculinity	0.07	0.11	0.07	0.13	0.07	0.11	0.05	0.26	0.06	0.14
Uncertainty avoidance	0.27	0.001	0.28	0.001	0.22	0.001	0.27	0.001	0.09	0.03
Long-term orientation	0.08	0.001	0.04	0.38	0.07	0.11	0.07	0.11	0.12	0.01
Indulgence	0.30	0.001	0.26	0.001	0.27	0.001	0.26	0.001	0.01	0.74
Individualism	0.35	0.001	0.33	0.001	0.36	0.001	0.37	0.001	0.02	0.73

1. GBV content and all categories of GBV content show a positive correlation with culture and all its parameters by both the correlation metrics. For example, the correlation between GBV content and culture is 0.55, with a significant *p*-value of 0.001. Similarly, the correlation between culture and other categories of GBV *i.e.,* sexual violence, physical violence, and harmful practices content is 0.51, 0.53, and 0.53, respectively with significant *p*-values.
2. The three parameters of culture *uncertainty avoidance*, *indulgence*, and *individualism* show comparatively higher correlation values as compared to other parameters of culture *power distance*, *masculinity*, and *long-term orientation*. This observation is consistent with all the categories of GBV content and with both the correlation metrics. The Pearson correlation of *uncertainty avoidance*, *indulgence*, and *individualism* with GBV tweet content is 0.36, 0.34, and 0.45 (Table 1). On the other hand, the pearson correlation of *power distance*, *masculinity*, and *long-term orientation* with GBV tweets content is 0.27, 0.16, and 0.17 respectively.
3. We also observe that the same content analyses performed for GBV content did not show a similar strong and consistent correlation with generic content. The Pearson correlation between generic content and culture is 0.33, which is much lower than the Pearson correlation between GBV content and culture, *i.e.,* 0.55. Additionally, generic



(a) GBV-culture (b) Sexual-Culture (c) Physical-Culture

(d) Harmful-Culture (e) Generic-Culture

Figure 2 Scatter plots between different categories of content distance (A: GBV, B: sexual, C: physical, D: harmful practices, E: generic) and cultural distance.

Full-size DOI: 10.7717/peerjcs.1051/fig-2

content fails to show any correlation with culture and its parameters from spearman correlation.

Observation 1 indicates that GBV content has an influence of culture and all six parameters of culture. The observation is consistent with all the categories of GBV content, *i.e.,* sexual violence, physical violence, and harmful practices. Additionally, we also show the scatter plots of the culture and different GBV content types in Fig. 2 to reconfirm the findings. All these results imply that countries with similar culture also show higher similarity in GBV content. GBV content is composed of discussions, news, comments generated by users on the topic related to GBV. The reason for the similarity in GBV content for similar culture countries is the similarity in their discussions, news, and comments. For further understanding, we manually discerned the content of similar culture countries. According to Hofstede's dimensions, the USA and Canada are more similar[8] in culture as compared to USA and Iran. Similarly, Iran is more similar to Arab countries as compared to the USA. Table 6 shows the scores of cultural distance between different countries using Hofstede's dimensions. In order to show the content differences of different culture countries, we ploted the word clouds of common frequent words of USA-Canada and Arab countries-Iran in Figs. 3A and 3B, respectively. For the countries USA and Canada, we found keywords like *gfriend, whitesupremacist, objectifying* as common frequent keywords. For the countries Arab countries and Iran, we found keywords like *veiled, hijab, attacked, predator* as common frequent keywords.

The highlights in observation 2 suggest that a few parameters of culture also play an important role in shaping the content related to GBV. Interestingly, Hofstede's parameters

[8]This similarity is according to the Hofstede's dimensions calculated using Eq. (1)

Table 6 Cultural distance between different countries by Hofstede's dimensions.

	Arab Countries	Argentina	Australia	Brazil	Canada	China	Colombia	France	Germany	India	Indonesia	Iran	Italy	Japan	Philippines	Russia	Spain	Thailand	Turkey	UK	USA
Arab Countries	0	46.3	78.9	35.6	72.1	78.7	59.7	59	81.7	41.7	50.6	28.3	64.7	85.8	31.8	69.2	42.9	34	35.9	89.1	78.4
Argentina	46.3	0	58.3	34.2	54.8	104	44.9	56.6	74.4	71.5	75.6	38.9	62.7	80.9	67	89.1	36.9	47.7	35.9	75.8	61.7
Australia	78.9	58.3	0	71.4	20.5	117	88.4	71.8	74.4	79.8	102	64.9	66.1	101	86.4	120	70.3	85.3	78.1	34.5	7.9
Brazil	35.6	34.2	71.4	0	60	77.5	48.8	41.7	65.4	51.5	47.3	41.5	58.5	70.1	49.7	63.7	26.8	32.6	14.5	76.7	71.6
Canada	72.1	54.8	20.5	60	0	102	84.8	60.2	60.4	67.4	86.7	58	57.7	92.5	79.1	107	58.8	73.2	66.8	26.3	18.2
China	78.7	104	117	77.5	102	0	108	87	75.9	48.3	40.1	89.6	82.4	79.8	66.9	75.5	84.7	77.4	79.7	101	112
Colombia	59.7	44.9	88.4	48.8	84.8	108	0	87.3	104	87.5	76.9	59.7	97.5	98.3	65.4	104	69.4	54.8	56.3	102	91.4
France	59	56.6	71.8	41.7	60.2	87	87.3	0	50.1	59.3	70	65.4	39.1	64.8	75.7	53.6	28.1	64.8	38.6	71.9	70.6
Germany	81.7	74.4	74.4	65.4	60.4	75.9	104	50.1	0	63.9	76.1	81	31.5	49	90.7	79.8	54.9	81.9	64.6	56.9	70.8
India	41.7	71.5	79.8	51.5	67.4	48.3	87.5	59.3	63.9	0	39.7	50.3	55.3	79.8	37.7	68.8	55.1	52.3	54.3	73.8	75.4
Indonesia	50.6	75.6	102	47.3	86.7	40.1	76.9	70	76.1	39.7	0	60	77.4	81.4	46.1	62.1	59.2	40	49.4	95.7	99.3
Iran	28.3	38.9	64.9	41.5	58	89.6	59.7	65.4	81	50.3	60	0	68.4	96.7	47.3	87.1	44.7	30.6	43.1	78.7	65.3
Italy	64.7	62.7	66.1	58.5	57.7	82.4	97.5	39.1	31.5	55.3	77.4	68.4	0	51.7	78.6	72.6	44.3	76.6	56	60.8	63.1
Japan	85.8	80.9	101	70.1	92.5	79.8	98.3	64.8	49	79.8	81.4	96.7	51.7	0	93.4	74.7	67.1	91.9	68	91.8	100
Philippines	31.8	67	86.4	49.7	79.1	66.9	65.4	75.7	90.7	37.7	46.1	47.3	78.6	93.4	0	82.6	66.2	48.7	56.8	90.2	84.2
Russia	69.2	89.1	120	63.7	107	75.5	104	53.6	79.8	68.8	62.1	87.1	72.6	74.7	82.6	0	57.1	72.6	55.2	117	118
Spain	42.9	36.9	70.3	26.8	58.8	84.7	69.4	28.1	54.9	55.1	59.2	44.7	44.3	67.1	66.2	57.1	0	42.6	17.9	76.1	70.5
Thailand	34	47.7	85.3	32.6	73.2	77.4	54.8	64.8	81.9	52.3	40	30.6	76.6	91.9	48.7	72.6	42.6	0	32.6	91.8	85.4
Turkey	35.9	35.9	78.1	14.5	66.8	79.7	56.3	38.6	64.6	54.3	49.4	43.1	56	68	56.8	55.2	17.9	32.6	0	84	78.5
UK	89.1	75.8	34.5	76.7	26.3	101	102	71.9	56.9	73.8	95.7	78.7	60.8	91.8	90.2	117	76.1	91.8	84	0	28.5
USA	78.4	61.7	7.9	71.6	18.2	112	91.4	70.6	70.8	75.4	99.3	65.3	63.1	100	84.2	118	70.5	85.4	78.5	28.5	0



(a) USA-Canada (b) Arab countries-Iran

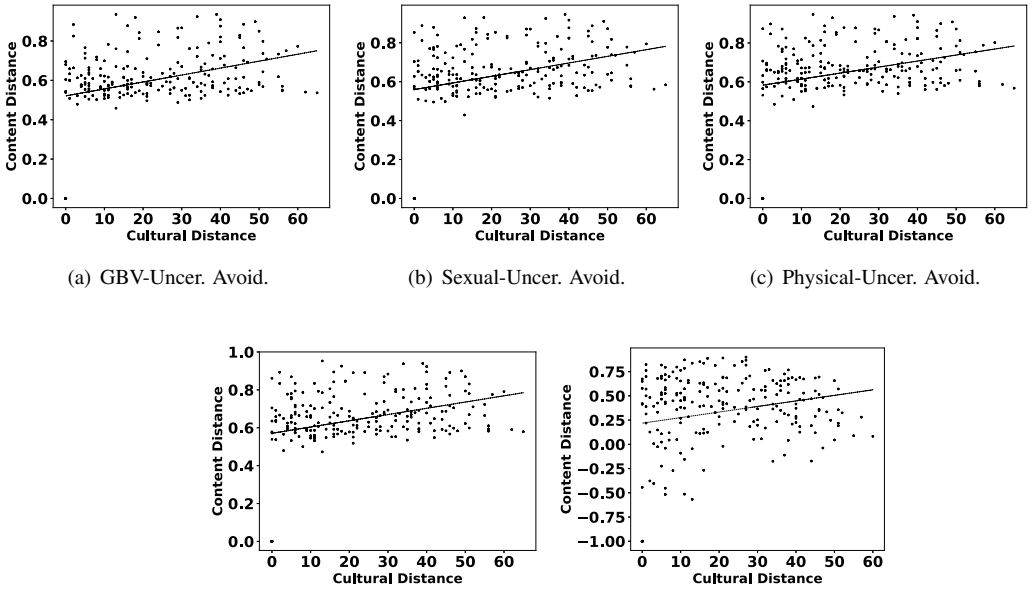
Figure 3 **Word Clouds showing common frequent keywords of culturally similar countries (A: USA-Canada and B: Arab Countries-Iran).**

Full-size DOI: 10.7717/peerjcs.1051/fig-3

uncertainty avoidance, indulgence, and individualism are found to show more impact on GBV related content than other parameters like power distance, masculinity, and long term orientation. For further reconfirming the connection between these parameters and different types of GBV content, we also show the scatter plots of these parameters and different content in Figs. 4, 5 and 6. From the scatter plots, we can evidently observe that Hofstede's parameters uncertainty avoidance, indulgence, and individualism consistently show a close association (points are closer to the fitted line) with all types of GBV content. The same is not true for generic content. This shows a connection between these parameters of culture and different GBV categories content. The reasons for more influence of these parameters require further exploration which is outside the scope of this work. However, this observation again recommends a role of culture on GBV.

Observation 3 further strengthens the findings of Observations 1 and 2. The pattern of correlation showing a connection between culture and different categories of GBV is not the same for generic content. The lower and inconsistent correlation values of the generic content as compared to GBV content reinforce a stronger relationship between GBV content and the culture of a country. Further, the scatter plots shown in Figs. 2, 4, 5 and 6 also show that the points in all the plots of generic content are more scattered from the fitted line as compared to points in GBV and its category content plots. For example, the points in the GBV-culture plot (Fig. 2A) are closer to the fitted line, while in the generic-culture plot (Fig. 2E), the points are farther to the fitted lines showing a comparatively lower correlation. Other categories of content (sexual, physical, and harmful) also show a stronger correlation with culture as compared to generic content.

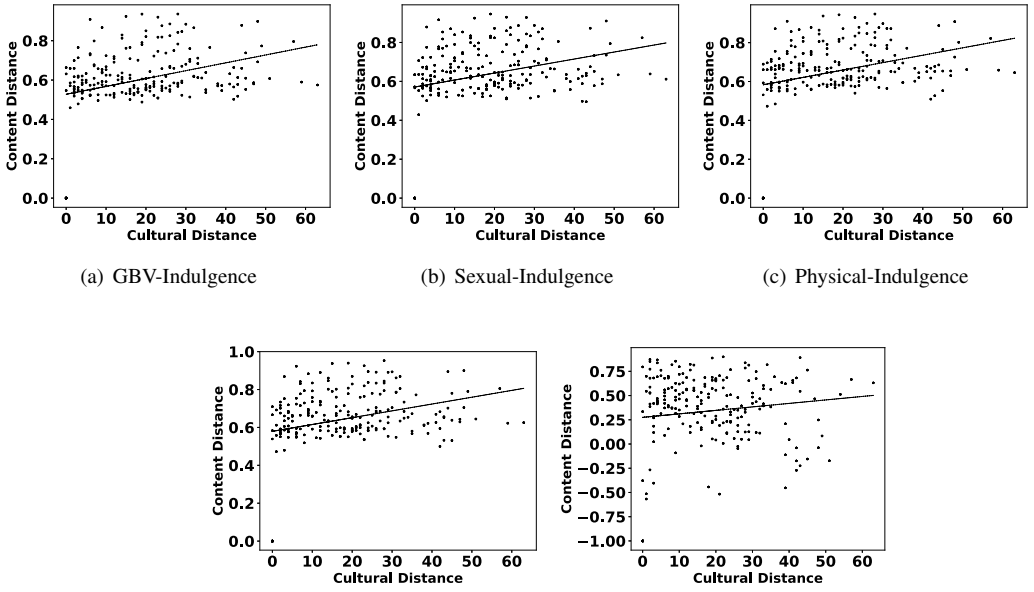
Generic content is composed of content from different topics, a few of which can be highly correlated to culture (*Cheke et al., 2020*), such as food, and a few can hardly show any correlation (*Annamoradnejad et al., 2019*), such as technology. These characteristics of generic content can be the most probable reason for showing weak correlations. Here we showed results of generic content just to provide a broader background for understanding. Next, we describe the results from graph analyses in order to validate findings from correlation analyses.



(a) GBV-Uncer. Avoid. (b) Sexual-Uncer. Avoid. (c) Physical-Uncer. Avoid.

(d) Harmful-Uncer. Avoid. (e) Generic-Uncer. Avoid.

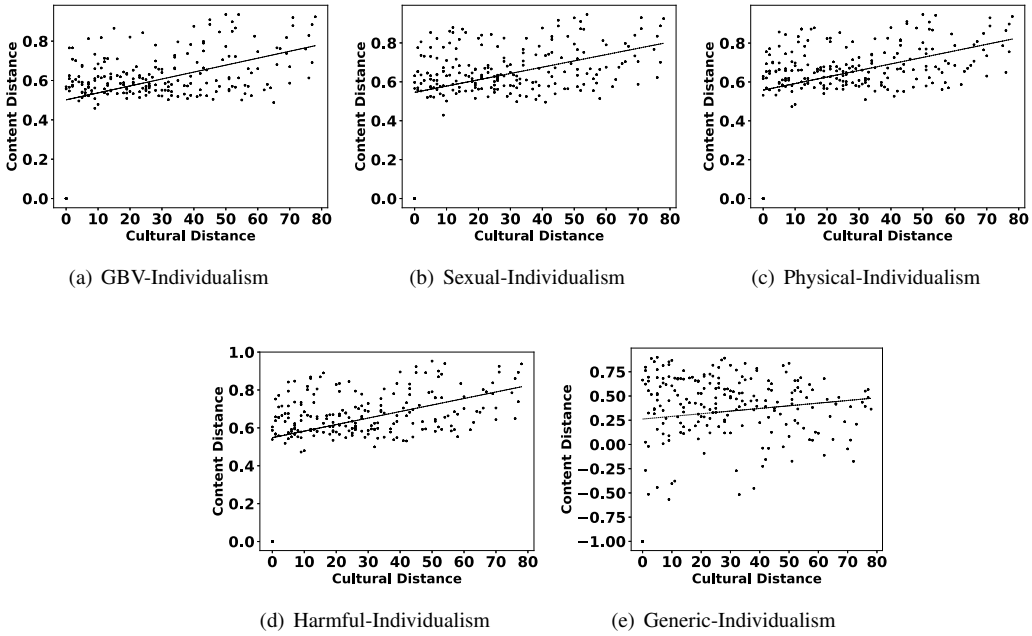
Figure 4 **Scatter plots between different categories of content distance (A: GBV, B: sexual, C: physical, D: harmful practices, E: generic) and cultural distance (uncertainty avoidance).**
Full-size DOI: 10.7717/peerjcs.1051/fig-4



(a) GBV-Indulgence (b) Sexual-Indulgence (c) Physical-Indulgence

(d) Harmful-Indulgence (e) Generic-Indulgence

Figure 5 **Scatter plots between different categories of content distance (A: GBV, B: sexual, C: physical, D: harmful practices, E: generic) and cultural distance (indulgence).**
Full-size DOI: 10.7717/peerjcs.1051/fig-5



(a) GBV-Individualism (b) Sexual-Individualism (c) Physical-Individualism

(d) Harmful-Individualism (e) Generic-Individualism

Figure 6 **Scatter plots between different categories of content distance (A: GBV, B: sexual, C: physical, D: harmful practices, E: generic) and cultural distance (individualism).**
Full-size DOI: 10.7717/peerjcs.1051/fig-6

Table 7 **Statistical inferences from the created graphs.**

Graph attributes	GBV	Culture	Sexual	Physical	Harmful	Generic
Edges	105	105	105	105	105	105
Nodes	21	21	21	21	21	21
Clustering coefficient	0.71	0.74	0.73	0.69	0.71	0.80
No of nodes in largest connected component	21	21	20	20	19	19
No of connected components	1	1	2	2	3	3

Graph analysis

We first summarize the statistical characteristics of the created graphs in our research in Table 7. All the graphs show similar basic properties because there is node correspondence in all the graphs. The edges and edge weights are the main varying parameter in the graphs. Next, we explain the details of graph comparison.

Table 8 shows the distance between different created graphs from various metrics. For each distance metric, if the distance value between a pair of graphs (G1, G2) is smaller than the distance between another pair of graphs (G3, G2), it means that the graph G2 is more similar to G1 as compared to G3. From the table, we observe that for all the metrics, the distance between the generic tweet graph and the culture graph is consistently higher than the distance between other graphs (GBV-culture, sexual violence-culture, physical violence-culture, and harmful practices-culture). For example, the metric *QuantumJSD*, the distance between generic graph and culture graph is 0.27 while for GBV graph and culture graph is 0.21. For the same metric, the distance between the sexual graph-culture

Table 8 Comparative results of distances between different graphs.

Distance metric	GBV-Culture	Sexual-Culture	Physical-Culture	Harmful-Culture	Generic-Culture
QuantumJSD	0.21	0.21	0.20	0.22	**0.27**
DegreeDivergence	0.19	0.26	0.34	0.37	**0.38**
JaccardDistance	0.70	0.70	0.72	0.71	**0.76**
Hamming	0.32	0.29	0.31	0.31	**0.41**
HammingIpsenMikhailov	0.25	0.24	0.25	0.26	**0.32**
Frobenius	12.16	11.66	12.0	12.0	**13.71**
NetLSD	2.71	11.62	11.18	11.97	**21.78**

graph, the physical graph-culture graph, and the harmful graph-culture graph is 0.21, 0.20, and 0.22, respectively. This shows that the graph created using content distance by GBV and its categories are more similar among themselves and to the graph created using cultural distance. The graph created using the generic content is consistently more distant from the cultural graph as compared to other graphs. This observation re-validates the observation from correlation analyses showing a higher degree of similarity between GBV content for similar culture countries. As two similar graphs show lesser distance by graph distance metrics, two similar graphs will also show similar clusters. Next, we present clustering in the created graphs.

Clustering the graphs

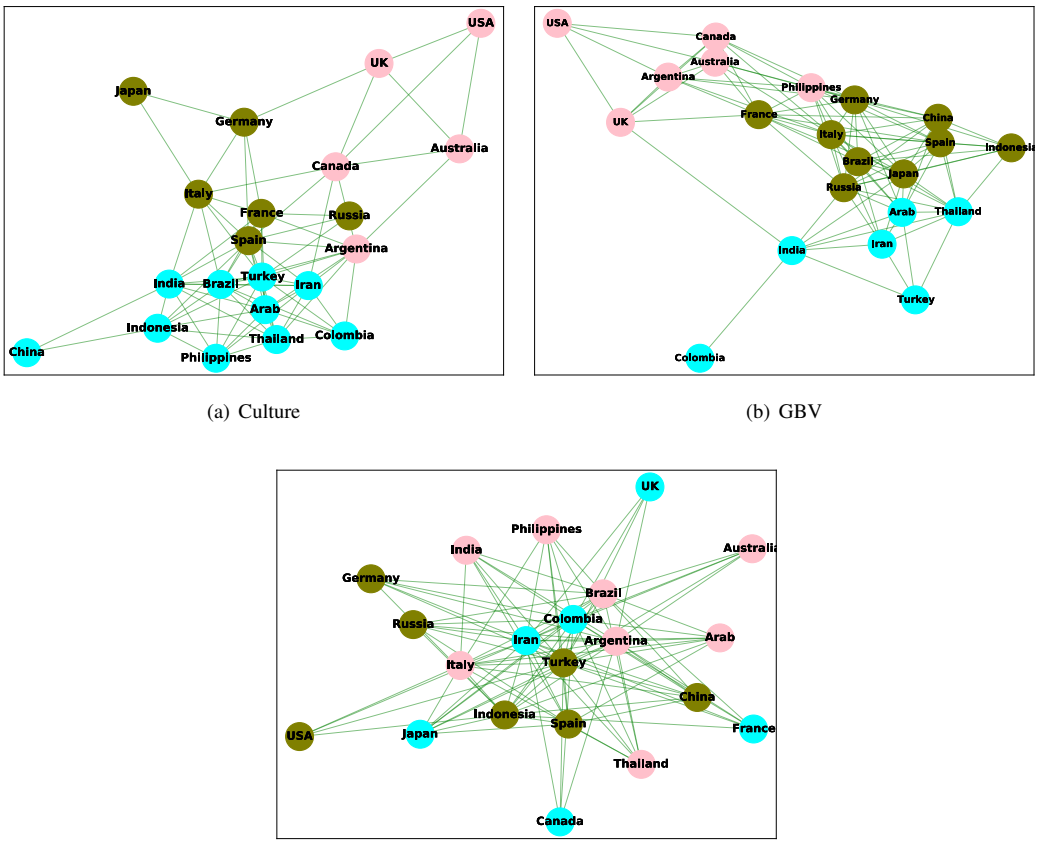
Figure 7 shows the plots of clusters in the culture graph, GBV graph, and generic content graph. From the created clusters, we observe that the clusters in the culture graph are more similar to GBV graph as compared to generic graph. The countries in the culture graph which belong to the same cluster have a larger overlap with the GBV graph rather than the generic graph. For example, the countries *USA, UK, and Australia* belong to the same cluster in the culture graph, and the same is also true for GBV graph. However, for the generic graph, all three countries belong to different clusters. This observation again shows a higher similarity between the culture graph and the GBV graph than the generic graph and the culture graph. Thus, we observe that the created clusters are also congruous to all other findings stating a higher level of relation between culture and GBV content, which is not the same for generic content.

DISCUSSIONS

Implications

In this research, we use social media data to verify connections between the culture of a country and GBV. Our findings suggest that real-world hypotheses are also evident in social media data, and their verification is no longer dependent on survey-based data. We believe that this research not only validates the hypothesis of confluence between culture and GBV but also points to the possibility of verification of other hypotheses related to GBV.

A finer analysis can also reveal culture-specific traits of GBV, which can further enhance understanding of GBV across cultures. We argue that these analyses are



(a) Culture

(b) GBV

(c) Generic

Figure 7 **Clustering nodes (countries) in the culture graph, GBV graph, and generic graph.**
Full-size DOI: 10.7717/peerjcs.1051/fig-7

vital for designing culture-aware policies and strategies to curb GBV. There is a huge possibility of discovery of many more cultural norms like those pronounced by the World Bank (https://thedocs.worldbank.org/en/doc/656271571686555789-0090022019/original/ShiftingCulturalNormstoAddressGBV.pdf), which can promote GBV. Thus, this research paves a path for understanding culture-specific GBV using online social network data.

Limitations and critiques

In this section, we show a few possible limitations and how our research overcomes those. In this article, we have performed cross-cultural research using online content from Twitter. Here, we limit our research to English tweets only pertaining to two reasons. First, the GBV keywords are in English, resulting in a collection of English GBV content. Second, English has become the new lingua franca on Twitter (*Choudhary et al., 2018*), which delivers sufficient tweets for this research.

GBV data collection is based upon GBV keywords provided by UNFPA, which is a global organization. The provided keywords can be incomplete and non-exhaustive. There might have scope for increasing these keywords; however, GBV is a sensitive issue, and extending

keywords without the intervention of social experts may introduce errors. So, we limit this research to globally available keywords only.

Online content is much inflected by a flux of ongoing news and events, which can lead to differences in data patterns in certain time periods. However, our research is based upon data from a long temporal span which diffuses such temporal inflections (*Grieve, Nini & Guo, 2018*).

There can be many more ways to capture the distance between countries in terms of GBV, but we have limited this to content distance using two common metrics (cosine similarity and jaccard similarity). The content distance used in this research captures the basic difference between tweet tokens of the two countries. However, the same methodology can be easily adapted to other twitter features and metrics.

CONCLUSION AND FUTURE WORK

The article investigates evidence of the confluence between culture and GBV with the help of social media content. Social media content is explored by means of correlation analyses and graph-based analyses to find the traces of culture in GBV related social media content. In this research, we find a noteworthy influence of culture on GBV related content which is not apparent in generic content. The observation is consistent with different analyses and metrics. This research not only claims higher confluence between GBV and culture but also paves a path for effective policy-making and research related to GBV by means of social media content. Social media content captures behavioral aspects related to GBV, which can be used for other investigations related to GBV. As a future work of this research, we would like to understand the role of other factors like economy, unemployment,and crises in GBV. Moreover, this research is a global analysis of different countries of the world. We would also like to extend the research to a finer scale of states or counties within a country.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Rimjhim conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Sourav Dandapat conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available at Zenodo: Rimjhim, & Sourav Dandapat. (2022). Is Gender Based Violence a Confluence of Culture? Let's look for empirical evidence via Social Media. https://doi.org/10.5281/zenodo.6820256.

REFERENCES

Abdullah S, Murnane EL, Costa JM, Choudhury T. 2015. Collective smile: measuring societal happiness from geolocated images. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. New York: ACM, 361–374.

Abrahams N, Devries K, Watts C, Pallitto C, Petzold M, Shamu S, GarcÍa-Moreno C. 2014. Worldwide prevalence of non-partner sexual violence: a systematic review. *The Lancet* **383(9929)**:1648–1654 DOI 10.1016/S0140-6736(13)62243-6.

Aldridge J. 2021. Not an either/or situation: the minimization of violence against women in United Kingdom domestic abuse policy. *Violence Against Women* **27(11)**:1823–1839 DOI 10.1177/1077801220927079.

Alesina A, Brioschi B, La Ferrara E. 2016. Violence against women: a cross-cultural analysis for Africa. *Economica* **88(349)**:70–104 DOI 10.1111/ecca.12343.

Annamoradnejad I, Fazli M, Habibi J, Tavakoli S. 2019. Cross-cultural studies using social networks data. *IEEE Transactions on Computational Social Systems* **6(4)**:627–636 DOI 10.1109/TCSS.2019.2919666.

Bahl T, Syed MH. 2003. *Encyclopaedia of Muslim world*. New Dehli: Anmol Publications.

Bishwajit G, Sarker S, Yaya S. 2016. Socio-cultural aspects of gender-based violence and its impacts on womens health in South Asia. *F1000Research* **5(802)**:802 DOI 10.12688/f1000research.8633.1.

Bisika T. 2008. Do social and cultural factors perpetuate gender based violence in Malawi? *Gender and Behaviour* **6(2)**:1884–1896.

Blake KR, ODean SM, Lian J, Denson TF. 2021. Misogynistic tweets correlate with violence against women. *Psychological Science* **32(3)**:315–325 DOI 10.1177/0956797620968529.

Cheke N, Chandra J, Dandapat SK. 2020. Understanding the impact of geographical distance on online discussions. *IEEE Transactions on Computational Social Systems* **7(4)**:858–872 DOI 10.1109/TCSS.2020.2993450.

Choudhary N, Singh R, Rao VA, Shrivastava M. 2018. Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In: *Proceedings of the 27th international conference on computational linguistics*. 1570–1577.

Chowdhury AG, Sawhney R, Mathur P, Mahata D, Shah RR. 2019. Speak up, fight back! detection of social media disclosures of sexual harassment. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: student research workshop*. 136–146.

Dartnall E, Jewkes R. 2013. Sexual violence against women: the scope of the problem. *Best Practice & Research Clinical Obstetrics & Gynaecology* **27(1)**:3–13 DOI 10.1016/j.bpobgyn.2012.08.002.

De Domenico M, Biamonte J. 2016. Spectral entropies as information-theoretic tools for complex network comparison. *Physical Review X* **6(4)**:041062.

De Meo P, Ferrara E, Fiumara G, Provetti A. 2011. Generalized louvain method for community detection in large networks. In: *2011 11th international conference on intelligent systems design and applications*. 88–93.

Djamba YK, Kimuna SR. 2015. *Gender-based violence: perspectives from Africa, the Middle East, and India*. Cham: Springer International Publishing.

Dredze M, Osborne M, Kambadur P. 2016. Geolocation for twitter: timing matters. In: *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies*. 1064–1069.

Elischberger HB, Glazier JJ, Hill ED, Verduzco-Baker L. 2018. Attitudes toward and beliefs about transgender youth: a cross-cultural comparison between the United States and India. *Sex Roles* **78(1)**:142–160 DOI 10.1007/s11199-017-0778-3.

ElSherief M, Belding E, Nguyen D. 2017. # notokay: understanding gender-based violence in social media. In: *Eleventh international AAAI conference on web and social media*.

Fakunmoju S, Bammeke F. 2017. Gender-based violence beliefs and stereotypes: cross-cultural comparison across three countries. *International Journal of Asian Social Science* **7(9)**:738–753 DOI 10.18488/journal.1.2017.79.738.753.

Fulu E, Miedema S. 2015. Violence against women. *Violence Against Women* **21(12)**:1431–1455 DOI 10.1177/1077801215596244.

García-Gavilanes R, Mejova Y, Quercia D. 2014. Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. In: *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing*. New York: ACM, 1511–1522.

Garcia-Gavilanes R, Quercia D, Jaimes A. 2013. Cultural dimensions in twitter: time, individualism and power. In: *Proceedings of the international AAAI conference on web and social media, volume 7(1)*.

Grieve J, Nini A, Guo D. 2018. Mapping lexical innovation on american social media. *Journal of English Linguistics* **46(4)**:293–319 DOI 10.1177/0075424218793191.

Guerrero L. 2013. Burns due to acid assaults in Bogotá, Colombia. *Burns* **39(5)**:1018–1023 DOI 10.1016/j.burns.2012.10.022.

Hansson K, Sveningsson M, Ganetz H. 2021. Organizing Safe Spaces:# MeToo Activism in Sweden. In: *Computer supported cooperative work (CSCW)*. 1–32.

Hassan N, Poudel A, Hale J, Hubacek C, Huq KT, Santu SKK, Ahmed SI. 2020. Towards automated sexual violence report tracking. In: *Proceedings of the international AAAI conference on web and social media, volume 14*. 250–259.

Hébert-Dufresne L, Grochow JA, Allard A. 2016. Multi-scale structure and topological anomaly detection via a new network statistic: the onion decomposition. *Scientific Reports* **6**:31708 DOI 10.1038/srep31708.

Herrero Á, Jiménez A, Alcalde R. 2021. Advanced feature selection to study the internationalization strategy of enterprises. *PeerJ Computer Science* **7**:e403 DOI 10.7717/peerj-cs.403.

Hofstede G, Hofstede GJ, Minkov M. 2005. *Cultures and organizations: software of the mind*. Vol. 2. New York: McGraw-Hill.

Huang V, Beshai S, Yu M. 2016. The effects of the gender-culture interaction on self-reports of depressive symptoms: cross-cultural study among Egyptians and Canadians. *PeerJ* **4**:e2783 DOI 10.7717/peerj.2783.

Jewkes R, Jama-Shai N, Sikweyiya Y. 2017. Enduring impact of conflict on mental health and gender-based violence perpetration in Bougainville, Papua New Guinea: a cross-sectional study. *PLOS ONE* **12**(**10**):e0186062 DOI 10.1371/journal.pone.0186062.

Jewkes R, Levin J, Penn-Kekana L. 2002. Risk factors for domestic violence: findings from a South African cross-sectional study. *Social Science & Medicine* **55**(**9**):1603–1617 DOI 10.1016/S0277-9536(01)00294-5.

Johnson SE, Hoffman JA, Kruger LJ, Rizzo CJ. 2015. Teaching school psychology students about teen dating violence: a snapshot of training practices across the USA. *School Mental Health* **7**(**4**):249–260 DOI 10.1007/s12310-015-9156-x.

Karuna P, Purohit H, Stabile B, Hattery A. 2016. On the dynamics of local to global campaigns for curbing gender-based violence. ArXiv preprint. arXiv:1608.01648.

Khatua A, Cambria E, Khatua A. 2018. Sounds of silence breakers: exploring sexual violence on twitter. In: *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. Piscataway: IEEE, 397–400.

Kim ME. 2021. Shifting the lens: an implementation study of a community-based and social network intervention to gender-based violence. *Violence Against Women* **27**(**2**):222–254 DOI 10.1177/1077801219889176.

Kosub S. 2019. A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters* **120**:36–38 DOI 10.1016/j.patrec.2018.12.007.

Kulshrestha J, Kooti F, Nikravesh A, Gummadi PK. 2012. Geographic dissection of the twitter network. In: *ICWSM*.

Le Mat ML, Kosar-Altinyelken H, Bos HM, Volman ML. 2019. Discussing culture and gender-based violence in comprehensive sexuality education in Ethiopia. *International Journal of Educational Development* **65**:207–215 DOI 10.1016/j.ijedudev.2018.08.004.

Liu Y, Li Q, Liu X, Zhang Q, Si L. 2019. Sexual harassment story classification and key information identification. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. 2385–2388.

Malagoli LG, Stancioli J, Ferreira CH, Vasconcelos M, Couto da Silva AP, Almeida JM. 2021. A look into COVID-19 vaccination debate on twitter. In: *13th ACM web science conference 2021*. New York: ACM, 225–233.

Marine S, Lewis R. 2020. *Collaborating for change: transforming cultures to end gender-based violence in higher education*. Oxford: Oxford University Press.

Martínez F, Pacheco C, Galicia M. 2021. The# MeToo movement in twitter: fighting gender-based violence. In: *International conference on information technology & systems*. 36–44.

McCabe S, Torres L, LaRock T, Haque SA, Yang C.-H., Hartle H, Klein B. 2020. netrd: a library for network reconstruction and graph distances. ArXiv preprint. arXiv:2010.16019.

Moitra A, Ahmed SI, Chandra P. 2021. Parsing the'Me'in# MeToo: sexual harassment, social media, and justice infrastructures. *Proceedings of the ACM on Human-Computer Interaction* **5(CSCW1)**:1–34.

Nikolov D, Oliveira DF, Flammini A, Menczer F. 2015. Measuring online social bubbles. *PeerJ Computer Science* **1**:e38 DOI 10.7717/peerj-cs.38.

Oggier F, Datta A. 2021. Renyi entropy driven hierarchical graph clustering. *PeerJ Computer Science* **7**:e366 DOI 10.7717/peerj-cs.366.

Pal J, Thawani U, Van Der Vlugt E, Out W, Chandra P. 2018. Speaking their mind: populist style and antagonistic messaging in the tweets of Donald Trump, Narendra Modi, Nigel Farage, and Geert Wilders. *Computer Supported Cooperative Work (CSCW)* **27(3)**:293–326 DOI 10.1007/s10606-018-9316-2.

Palmer E, Williams S. 2017. A shift in attitude? Institutional change and sexual and gender-based crimes at the Extraordinary Chambers in the Courts of Cambodia. *International Feminist Journal of Politics* **19(1)**:22–38 DOI 10.1080/14616742.2017.1283247.

Pandey R, Purohit H, Stabile B, Grant A. 2018. Distributional semantics approach to detect intent in twitter conversations on sexual assaults. In: *2018 IEEE/WIC/ACM international conference on web intelligence (WI)*. Piscataway: IEEE, 270–277.

Perrin N, Marsh M, Clough A, Desgroppes A, Phanuel CY, Abdi A, Kaburu F, Heitmann S, Yamashina M, Ross B. 2019. Social norms and beliefs about gender based violence scale: a measure for use with gender based violence prevention programs in low-resource and humanitarian settings. *Conflict and Health* **13(1)**:1–12 DOI 10.1186/s13031-018-0184-7.

Piedalue A, Gilbertson A, Alexeyeff K, Klein E. 2020. is gender-based violence a social norm? Rethinking power in a popular development intervention. *Feminist Review* **126(1)**:89–105 DOI 10.1177/0141778920944463.

Prakash CD, Majumdar A. 2021. Analyzing the role of national culture on content creation and user engagement on Twitter: the case of Indian Premier League cricket franchises. *International Journal of Information Management* **57**:102268 DOI 10.1016/j.ijinfomgt.2020.102268.

Puente SN, Maceiras SD, Romero DF. 2021. Twitter activism and ethical witnessing: possibilities and challenges of feminist politics against gender-based violence. *Social Science Computer Review* **39(2)**:295–311 DOI 10.1177/0894439319864898.

Purohit H, Banerjee T, Hampton A, Shalin VL, Bhandutia N, Sheth AP. 2015. Gender-based violence in 140 characters or fewer: A# BigData case study of Twitter. ArXiv preprint. arXiv:1503.02086.

Raj A, Silverman J. 2002. Violence against immigrant women: the roles of culture, context, and legal immigrant status on intimate partner violence. *Violence Against Women* **8(3)**:367–398 DOI 10.1177/10778010222183107.

Razi A. 2020. Deploying human-centered machine learning to improve adolescent online sexual risk detection algorithms. In: *Companion of the 2020 ACM international conference on supporting group work*. 157–161.

Samani ZR, Guntuku SC, Moghaddam ME, Preoţiuc-Pietro D, Ungar LH. 2018. Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr. *PLOS ONE* **13(7)**:e0198660 DOI 10.1371/journal.pone.0198660.

Sazzed S. 2021. Identifying vulgarity in Bengali social media textual content. *PeerJ Computer Science* **7**:e665 DOI 10.7717/peerj-cs.665.

Shin D, Chotiyaputta V, Zaid B. 2022. The effects of cultural dimensions on algorithmic news: How do cultural value orientations affect how people perceive algorithms? *Computers in Human Behavior* **126**:107007 DOI 10.1016/j.chb.2021.107007.

Silva TH, Viana AC, Benevenuto F, Villas L, Salles J, Loureiro A, Quercia D. 2019. Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)* **52(1)**:1–39.

Spencer-Oatey H, Franklin P. 2012. What is culture. In: *A compilation of quotations. GlobalPAD core concepts*. 1–22.

Starkey JC, Koerber A, Sternadori M, Pitchford B. 2019. # MeToo goes global: media framing of silence breakers in four national settings. *Journal of Communication Inquiry* **43(4)**:437–461 DOI 10.1177/0196859919865254.

Stewart LG, Arif A, Nied AC, Spiro ES, Starbird K. 2017. Drawing the lines of contention: networked frame contests within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* **1(CSCW)**:1–23.

Stubbs-Richardson M, Rader NE, Cosby A. G.. 2018. Tweeting rape culture: examining portrayals of victim blaming in discussions of sexual assault cases on Twitter. *Feminism & Psychology* **28(1)**:90–108 DOI 10.1177/0959353517715874.

Tantardini M, Ieva F, Tajoli L, Piccardi C. 2019. Comparing methods for comparing networks. *Scientific Reports* **9(1)**:1–19 DOI 10.1038/s41598-018-37186-2.

Wang Y, Mohd Pozi MS, Kawai Y, Jatowt A, Akiyama T. 2017. Exploring cross-cultural crowd sentiments on Twitter. In: *Proceedings of the 28th ACM conference on hypertext and social media*. New York: ACM, 321–322.

PeerJ Computer Science

Investigating toxicity changes of cross-community redditors from 2 billion posts and comments

Hind Almerekhi[1], Haewoon Kwak[2] and Bernard J. Jansen[3]

[1] Hamad Bin Khalifa University, Doha, Qatar
[2] Singapore Management University, Singapore, Singapore
[3] Qatar Computing Research Institute, HBKU, Doha, Qatar

ABSTRACT

This research investigates changes in online behavior of users who publish in multiple communities on Reddit by measuring their toxicity at two levels. With the aid of crowdsourcing, we built a labeled dataset of 10,083 Reddit comments, then used the dataset to train and fine-tune a Bidirectional Encoder Representations from Transformers (BERT) neural network model. The model predicted the toxicity levels of 87,376,912 posts from 577,835 users and 2,205,581,786 comments from 890,913 users on Reddit over 16 years, from 2005 to 2020. This study utilized the toxicity levels of user content to identify toxicity changes by the user within the same community, across multiple communities, and over time. As for the toxicity detection performance, the BERT model achieved a 91.27% classification accuracy and an area under the receiver operating characteristic curve (AUC) score of 0.963 and outperformed several baseline machine learning and neural network models. The user behavior toxicity analysis showed that 16.11% of users publish toxic posts, and 13.28% of users publish toxic comments. However, results showed that 30.68% of users publishing posts and 81.67% of users publishing comments exhibit changes in their toxicity across different communities, indicating that users adapt their behavior to the communities' norms. Furthermore, time series analysis with the Granger causality test of the volume of links and toxicity in user content showed that toxic comments are Granger caused by links in comments.

Subjects Data Mining and Machine Learning, Data Science, Network Science and Online Social Networks, Text Mining, Neural Networks
Keywords Reddit, Toxicity, Posting behavior, Online communities, Machine learning, Online hate

Submitted 2 May 2022
Accepted 18 July 2022
Published 18 August 2022

Corresponding author
Hind Almerekhi,
h.almerekhi@gmail.com,
hialmerekhi@hbku.edu.qa

Academic editor
Arkaitz Zubiaga

Additional Information and Declarations can be found on page 35

DOI **10.7717/peerj-cs.1059**

Copyright
2022 Almerekhi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Online social media platforms enable users to communicate with each other in various ways, like sharing and publishing different types of content (*Mondal, Silva & Benevenuto, 2017*). Unfortunately, the rapid growth of online communication on social media platforms has caused an explosion of malicious content in the form of harassment, profanity, and cyberbullying (*Hu et al., 2013*). A survey by Pew Research Center (*Vogels, 2020*) showed that 41% out of 10,093 American adults were personally harassed online, and 25% experienced severe forms of harassment. Moreover, 55% of the survey participants considered online

harassment a major problem. This concern was also shared by online moderators (*Cheng, Danescu-Niculescu-Mizil & Leskovec, 2015*), noting that posts and comments on many social media platforms can easily take a dark turn and become *toxic*. Therefore, there is a need for solutions that identify toxic content and limit its presence on online platforms.

One challenge with studying online toxicity is the multitude of forms it takes (*Davidson et al., 2017*). These forms include hate speech, which refers to offensive content that targets a specific trait in a group of people (*Silva et al., 2016*); harassment, which occurs when a user deliberately aggravates other users online (*Cheng, Danescu-Niculescu-Mizil & Leskovec, 2015*); and cyberbullying, which means targeting and intimidating victims through online communication (*Bowler, Knobel & Mattern, 2015*).

The previous classifications of toxicity forms show that toxic content often contains insults, threats, and offensive language, which, in turn, contaminate online platforms (*Mondal, Silva & Benevenuto, 2017*) by preventing users from engaging in discussions or pushing them to leave (*Newell et al., 2016*). Thus, several online platforms have implemented prevention mechanisms, such as blocklists (*Jhaver et al., 2018*), that block specific accounts from interacting with users. Other approaches to preventing toxicity include deploying human moderators and bots to remove toxic content (*Chandrasekharan et al., 2018*). These efforts, however, are not scalable enough to curtail the rapid growth of toxic content on online platforms (*Davidson et al., 2017*). There is also the psychological distress associated with exposing human moderators to firsthand accounts of toxicity (*Rodriguez & Rojas-Galeano, 2018*). These challenges call for developing effective automatic or semiautomatic solutions to detect toxicity from a large stream of content on online platforms.

Users of social media platforms have various reasons for spreading harmful content, like personal or social gain (*Squicciarini, Dupont & Chen, 2014*). Studies show that publishing toxic content (*i.e.,* toxic behavior) is contagious (*Tsikerdekis & Zeadally, 2014*; *Rodriguez & Rojas-Galeano, 2018*); the malicious behavior of users can influence non-malicious users and leads them to misbehave, which affects the overall well-being of online communities. As an example of toxic behavior (*Alfonso & Morris, 2013*), one Reddit user named *Violentacrez* created several communities on controversial topics such as gore, and his followers mimicked this behavior by creating communities with highly offensive content as well. This influence-based culture (*Johnson, 2018*) that users follow in online communities motivates studies like the current study to investigate the problem of toxic online posting and commenting behavior. Fueled by cases reported in *Alfonso & Morris (2013)*, this study focuses on the toxic behavior of users on Reddit. In particular, this research investigates the toxic cross-community behavior of users, which refers to publishing toxic content in more than one community.

This study argues that the toxicity of users' content may *change* based on the environment (*i.e.,* community) in which they participate. Therefore, the focus is to investigate changes in toxicity in two types of content that Reddit describes as follows:

- *Post*: is the top-level submission of a user that can be either a post, link, video, image, or poll.

- *Comment*: is the response of another user or the poster to a post or a comment.

This study uses an extensive collection of more than 2.293 billion published content, including 87 million posts and 2.205 billion comments, from more than 1.2 million unique users who published content in more than 107,000 unique subreddits from June 2005 to April 2020.

LITERATURE REVIEW

Studies of online toxicity typically tackle the problem of hate from three main perspectives: (1) toxic behavior characterization, (2) toxic behavior detection, and (3) toxic behavior in online communities.

Toxic behavior characterization

Investigating human behavior is essential for organizations that rely on users to drive business, understand the dynamics of online communities, and prevent hate (*Mathew et al., 2020*; *Yin & Zubiaga, 2021*). Negative behavior of humans in online spaces involves a lack of inhibition, including online aggressiveness that would not exist in similar situations offline (*Lapidot-Lefler & Barak, 2012*). *Suler (2004)* introduced the phrase ”toxic disinhibition” and defined it as the inhibition loss of users who act violently online, which holds no benefits and leads users to violate conventional coexistence rules (*Suler, 2004*). A typical form of toxic disinhibition is flaming behavior, which involves using hostile expressions to refer to other users in online communication. Textual features of flaming behavior include harsh language, negative connotations, sexual harassment, and disrespectful expressions (*Pelicon et al., 2021*). The definition of toxic disinhibition, or toxic behavior, varies based on the users, the communities, and the types of interactions (*Shores et al., 2014*). For instance, toxic behavior can consist of cyberbullying and deviance between players in massively multiplayer online games (MMOGs) (*Shores et al., 2014*; *Kordyaka, Jahn & Niehaves, 2020*) or incivility between social media platform users (*Maity et al., 2018*; *Pronoza et al., 2021*), among other scenarios. In this work, we define *toxic behavior* in online communities as disseminating (*i.e.,* posting) toxic content with hateful, insulting, threatening, racist, bullying, and vulgar language (*Mohan et al., 2017*).

Toxic behavior detection

There are two known methods for detecting toxic behavior on online platforms. The first relies on social network analysis (SNA) (*Wang & Lee, 2021*). The study of *Singh, Thapar & Bagga (2020)* exemplifies SNA usage and content-based analysis to detect cyberbullying (a form of toxic behavior) on Twitter. The study investigates the Momo Challenge, a fake challenge spread on Facebook and other social media platforms to entice younger users to commit violent acts. Researchers collected incidents related to the challenge by tracking 5,615 users' network graphs and 7,384 tweets using the Momo Challenge hashtag and relevant keywords. Findings showed that a small number of users employed keywords related to the Momo Challenge to cause cyberbullying events, whereas the majority used the keywords to warn other users about the dangerous challenge. Techniques involving SNA are suitable for detecting toxic behavior patterns and targeted attacks like cyberbullying.

However, these techniques must analyze the involved users' social profiles or relations to detect toxic behavior, which may not be available on platforms other than social media websites. Therefore, the second and most common method avoids this limitation by detecting toxic behavior in user-generated content (*Djuric et al., 2015*).

Analyzing user-generated content involves detecting toxicity; this is a heavily investigated problem (*Davidson et al., 2017*; *Ashraf, Zubiaga & Gelbukh, 2021*; *Obadimu et al., 2021*). To detect toxic content, some studies (*Nobata et al., 2016*) build machine learning models that combine various semantic and syntactic features. At the same time, other studies use deep multitask learning (MTL) neural networks with word2vec and pretrained GloVe embedding features (*Kapil & Ekbal, 2020*; *Sazzed, 2021*). As for open-source solutions, Google offers the Perspective API (*Georgakopoulos et al., 2018*; *Mittos et al., 2020*), which allows users to score comments based on their perceived toxicity (*Carton, Mei & Resnick, 2020*). The API uses pretrained machine learning models on crowdsourced labels to identify toxicity and improve online conversations (*Perspective, 2017*).

By using the outcomes of previous studies (*Wulczyn, Thain & Dixon, 2017*; *Georgakopoulos et al., 2018*), this work evaluates the performance of classical machine learning models (*Davidson et al., 2017*) and neural network models (*Del Vigna et al., 2017*) to detect toxicity at two levels from user content.

Toxic behavior in online communities

Online platforms continuously strive to improve user engagement through various forms of interaction. Websites such as Facebook and Reddit offer users the freedom to create communities of their own volition to interact with similar groups of users (*Johnson, 2018*). Despite the great interest in promoting healthy interactions among users in online communities, platforms struggle with the toxic behavior of some unsolicited users (*Shen & Rose, 2019*). This problem was evident on Reddit (*Almerekhi, Kwak & Jansen, 2020*; *Massanari, 2017*), where *Chandrasekharan et al. (2017a)* found that some of the topics discussed by communities were incredibly toxic, leading to the 2015 ban of two hateful communities due to users' fears that these groups would infect other communities. The study found that the ban successfully prevented hate groups from spreading their toxicity to other communities. Nevertheless, this ban broke one of Reddit's core self-moderation policies, which exasperated users who sought total freedom on Reddit.

In a similar vein, *Mohan et al. (2017)* investigated the impact of toxic behavior on the health of communities. The study defines health as user engagement relative to community size and measures toxicity with a commercial lexicon from Community Sift to filter toxic words and phrases. By analyzing 180 communities, the study found a high negative correlation between community health and toxicity. Additionally, the study showed that communities require stable toxicity levels to grow in size without declining health. Despite these findings, the study did not consider users when investigating toxicity and viewed communities through content, not content creators.

As for cases in which toxic behavior arises between communities on different platforms, *Chandrasekharan et al. (2017b)* proposed a solution that relies on building a Bag of Communities (BoC). The research identified the abusive behavior of users in nine

communities from Reddit, 4chan, MetaFilter, and Voat. By computing cross-platform post similarity, the proposed model achieved 75% accuracy without any training data from the target platform. Moreover, the BoC model can achieve an accuracy of 91% after seeing 100,000 human-moderated posts, which outperforms other domain-specific approaches. However, the study focused on cross-platform abusive behavior through content analysis without accounting for the users who behaved in an abusive or toxic manner.

Research questions

Given the literature review discussed earlier, this study aims to answer the following research questions:

RQ1: *How can the toxicity levels of users' content and users across different communities be detected?*

RQ2: *Does the toxicity of users' behavior change (a) across different communities or (b) within the same community?*

RQ3: *Does the toxicity of users change over time across different communities?*

METHODOLOGY

In our study, investigating toxic behavior on Reddit requires a rigorous process that starts with obtaining the corpus from Reddit to detect toxicity and ends with finding insights from users' behavior.

Data collection site

Reddit is an online community with over 2.8 million sub-communities (https://www.oberlo.com/blog/reddit-statistics; retrieved on 25 May 2022) that cover various topics from news to entertainment, incorporating a mix of cultures (*Massanari, 2017*). Those sub-communities are commonly known as "subreddits", denoted with the prefix "r/". The main activities that registered users (often called Redditors) perform include (a) creating subreddits, (b) submitting posts (*i.e.,* sharing content in the community), (c) commenting on the posts of others, and (d) replying to comments in discussion threads (*Choi et al., 2015*; *Kumar et al., 2018*).

Figure 1 shows a post along with responses (*i.e.,* comments) from "r/science." We can see in the figure that despite the strict moderation in this community, comments like "Statistically, pretty people are stupid" might be perceived by some Redditors as toxic.

Obtain corpus

In this study, we targeted users in the top 100 subreddits (ranked by subscriber count). These subreddits account for a major proportion of Reddit's overall activity because they attract the largest number of active users (*Choi et al., 2015*). First, we compiled a list of subreddits using two popular external services: RedditList (http://redditlist.com; retrieved on Aug. 29, 2017), and RedditMetrics (https://frontpagemetrics.com/top; retrieved on Aug. 29, 2017). We used both websites to curate a list of the top 100 largest safe-for-work subreddits based on subscriber count. In Appendix A, Tables A1 and A2 show the top 100 subreddits sorted by total subscribers. Note that while the subreddit r/announcements

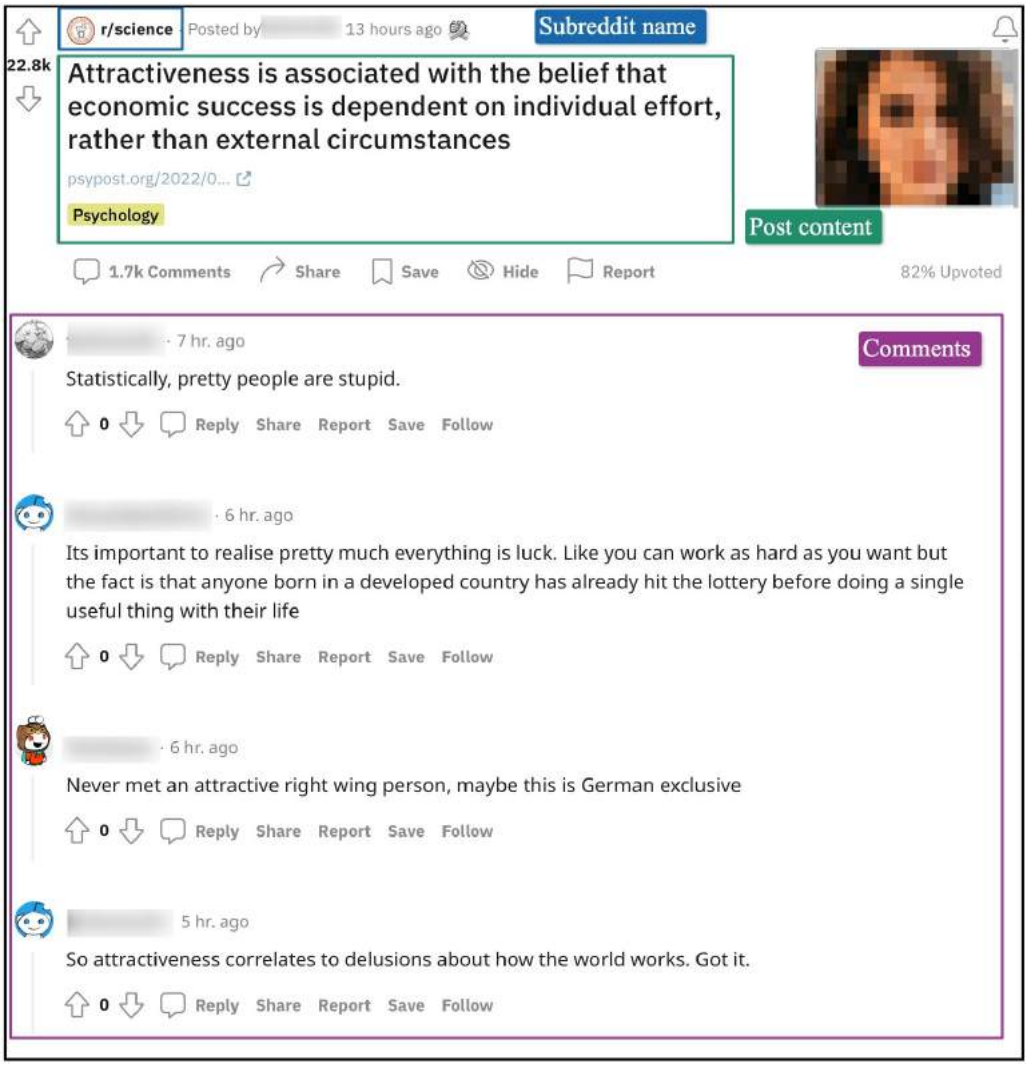


Figure 1 **A Reddit post from the subreddit "r/science" with its associated discussion threads.**
Full-size DOI: 10.7717/peerjcs.1059/fig-1

holds the highest number of subscribers, we removed it from our list because it did not serve our study purpose, as most of the users of the subreddit use it to consume content. Moreover, the subscriber values were retrieved on Aug. 29, 2017, while the total posts and comments are from June 2005 to April 2020. While there might be some differences among the top 100 subreddit rankings over the years, we used the list from 2017 because our study aims to understand the toxic behavior of a subgroup of users across multiple communities over time. Therefore, this specific list does not harm the research goal. Instead, the list brings unique opportunities for tracking the toxic behavior of a user subgroup over time.

Since this study focuses on users and their content, we cleaned our user collection by dropping any deleted or removed users (*i.e.*, users with a removed or deleted "author" field). This process yielded 3,208,002 users who posted and 5,036,095 users who commented from 2005 through August 2017 within the top 100 subreddits.

Additionally, we excluded bot users (*i.e.,* automated accounts) to avoid potential biases in the subsequent analysis by using a publicly available list of 320 bots on Reddit (https://www.reddit.com/r/autowikibot/wiki/redditbots; retrieved on May 13, 2019). Since the available bot list is outdated, it potentially misses newer bot accounts. Thus, for this work, we used the Pushshift API (https://pushshift.io; retrieved on May 22, 2019) to retrieve a list of accounts with a minimal comment reply delay. Setting the comment reply delay to 30 s allowed us to find more bot accounts that quickly reply to other users. We removed additional bot accounts by combining the bot list and Pushshift API list. When conducting this study, we found 37 bot accounts that produce around 2% of automated content. The massive volume of bot-generated content reaffirms the importance of removing bots in the data-cleaning phase of this study.

Since our research focuses on the toxic cross-community behavior of users, each user must participate in at least two different subreddits from the list of top 100 subreddits. Therefore, we filtered our original list of users to remove users who only participate in a single subreddit. This filtering process returned 577,835 users who posted (18% of the 3,208,002 users) and 890,913 users who left comments (17.7% of the 5,036,095). Furthermore, the intersection of these user lists yielded 241,138 users that performed both acts of posting and commenting. Overall, our dataset has 1,227,610 unique users who post and comment on Reddit. Lastly, we built our final collection of posts and comments with the Pushshift API (*Baumgartner, 2017*) by extracting user content from all subreddits. In other words, we started with a group of users participating in multiple communities, extracted their content from the top 100 subreddits, and then extracted their content from all other subreddits. As a result, we extracted 87,376,912 posts from 76,650 subreddits and 2,205,581,786 comments from 79,076 subreddits. To summarize, our collection has 2,292,958,698 posts and comments from 107,394 unique subreddits made by a group of cross-community users from June 2005 to April 2020.

Figure 2 shows the Cumulative Distribution Function (CDF) of the number of subreddits where a user left posts (A) and comments (B). Once a user participates in multiple subreddits (we already removed users who participated in a single subreddit, which is around 80% of users), the number of subreddits they participate in quickly grows. Findings from Fig. 2 indicate that users who participate in less than 10 subreddits are more than 80% and 50% in terms of posting and commenting, respectively. Also, participation through commenting seemed to be easier than posting; in Fig. 2, users who left comments in more than 20 subreddits are higher than 20%. In summary, the user collection in this study captured a substantial number of cross-community interactions and thus was appropriate for examining toxic behavior across multiple communities.

Label dataset

To investigate the toxicity of users, we required a reliable machine learning model to detect the toxicity of user content. However, before building the detection model, we first created a set of relevance judgments (*i.e.*, labels) that determine if a particular comment is toxic or not. Before conducting this study, we found a few publicly available toxicity detection datasets, such as the Wikipedia comments training set that Google's Jigsaw/Conversation AI released

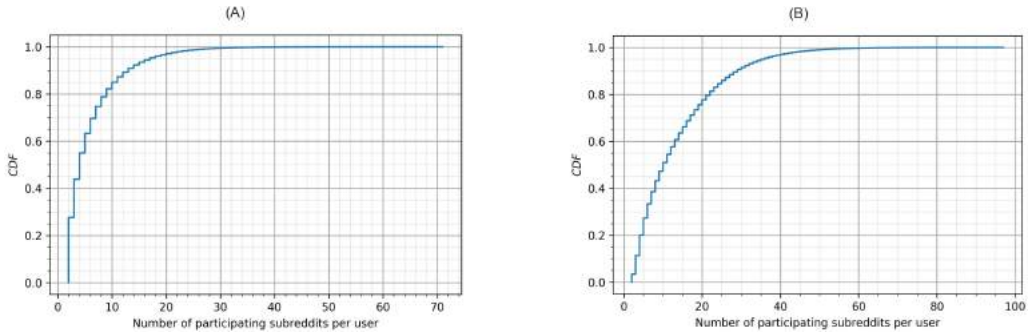


Figure 2 **Cumulative distribution function of the participating subreddits per user in (A) posts and (B) comments.**

Full-size DOI: 10.7717/peerjcs.1059/fig-2

on Kaggle (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge; retrieved on Feb. 13, 2018). However, we found that the dataset targets a classification problem that does not align well with our problem. Therefore, we created a labeled collection with data from our collection of comments.

To conduct our labeling experiment, we randomly sampled 10,100 comments from r/AskReddit, one of the largest subreddits in our collection. First, we used 100 comments to conduct a pilot study, after which we made minor modifications to the labeling task. Then, we proceeded with the remaining 10,000 comments to conduct the complete labeling task. We selected 10,000 comments to ensure that we had both a reasonably-sized labeled collection for prediction experiments and a manageable labeling job for crowdsourcing. For labeling, we recruited crowd workers from Appen (https://appen.com; retrieved on Jun. 10, 2022) (formerly known as Figure Eight). Appen is a widely used crowdsourcing platform; it enables customers to control the quality of the obtained labels from labelers based on their past jobs. In addition to the various means of conducting controlled experiments, this quality control makes Appen a favorable choice compared to other crowdsourcing platforms.

We designed a labeling job by asking workers to label a given comment as either toxic or nontoxic according to the definition of a toxic comment in the Perspective API (*Perspective, 2017*). If a comment was toxic, we asked annotators to rate its toxicity on a scale of two, as either (1) slightly toxic or (2) highly toxic. To avoid introducing any bias to the labeling task, we intentionally avoided defining what we consider highly toxic and slightly toxic and relied only on crowd workers' judgment on what the majority of annotators perceive as the correct label (*Vaidya, Mai & Ning, 2020*; *Hanu, Thewlis & Haco, 2021*). Nonetheless, we understand that toxicity is highly subjective, and different groups of workers might have varying opinions on what is considered highly or slightly toxic (*Zhao, Zhang & Hopfgartner, 2022*). Therefore, annotators had to pass a test by answering eight test questions before labeling to ensure the quality of their work.

Additionally, we used 70 test questions for quality control during the labeling process. The test questions for the labeling task were comments that were undoubtedly toxic or nontoxic. Such questions were given to annotators at random intervals to detect spammers

and ensure the quality of the obtained labels. One of the labeling task settings was to get the agreement of three annotators for each comment, meaning that at least two of the three annotators had to agree for a comment to be labeled. Moreover, crowd workers were required to spend a minimum of 30 s on each page while labeling and had a minimum labeling accuracy of 80% based on prior experience with Appen. Unfortunately, Appen does not provide any demographic statistics about the participating workers. Therefore, we only know that our workers had an accuracy $\geq$ 80% and could understand English.

To assist workers with the labeling task, we provided a link to the discussion thread of the comment on Reddit to establish the context of the discussion. Regarding the job design, we split the job into 2,000 pages, each with five rows. As for costs, we paid USD 490.2 for the pilot (USD 9.48) and the complete labeling task (USD 471.24). The details of the labeling job are in Appendix B, where we show the labeling job instructions in Fig. B1 and the labeling job test questions in Fig. B2.

To assess the quality of the obtained labels from crowdsourcing, we computed a set of reliability measures (http://mreliability.jmgirard.com; retrieved on Dec. 15, 2017) based on the labels from every worker for all classes. In this context, we define reliability as the proportion of perceived non-chance agreement to potential non-chance agreement (*Gwet, 2014*). To compute reliability, we used Eq. (1), where (P_o) is the percent of observed agreement and (P_c) is the percent of chance agreement. Generally, there are three main approaches to estimate the agreement by chance between two or more annotators across bifurcated categories: (a) individual-distribution-based approach like Cohen's Kappa (*Zhao, Liu & Deng, 2013*), (b) category-based approach such as Bennett et al.'s S score (*Bennett, Alpert & Goldstein, 1954*), and (c) average-distribution-based approaches, such as the ones adopted by Gwet's gamma coefficient (*Gwet, 2014*).

$$Reliability = \frac{P_\circ - P_c}{1 - P_c}. \tag{1}$$

Some of the reliability measures and results are in Table 1. Each measure takes into account various parameters based on its specific implementation. It is worth noting that while the observed agreement is high, Fleiss's Kappa is relatively low; this problem was heavily discussed by *Feinstein & Cicchetti (1990)* and *Salminen et al. (2019)*, with the low Kappa score attributed to the class imbalance. On the other hand, Gwet's gamma (*Gwet, 2014*) and other average distribution approaches handle such class imbalance differently and outperform the Kappa statistic measurements. Given the class imbalance in our dataset, representative of the online discussion community, we use Gwet's measure in our research. Lastly, we sampled 100 random comments from our training set and manually labeled them for their toxicity. Then, we measured the agreement between our labels and crowd workers. Our findings showed that there is an agreement of 92%, which means that crowd labels are suitable for training machine learning models.

Out of the 10,100 labeled comments, we removed three duplicate comments and 14 comments modified by moderators (*i.e.,* the actual content of the comments was either removed or heavily edited by automatic bots or moderators). Of the 10,083 labeled comments, 86.81% were labeled as nontoxic, while the remaining 13.82% were labeled

Table 1 Various reliability measure results on the relevance judgments obtained from Appen. ICC is Intraclass Correlation Coefficient.

Measure	Obtained value
Observed agreement	0.848
Gwet's gamma coefficientcitep (*Gwet, 2014*)	0.697
Average rater consistency ICC	0.738
Average rater agreement ICC	0.737
Bennett et al. Score (*Bennett, Alpert & Goldstein, 1954*)	0.697
Conger's kappa coefficient	0.483
Scott's pi, Krippendorff's alpha & Fleiss's kappa coefficient	0.482

Table 2 Examples of comments from every class as labeled by Appen workers.

Comment	Class
...Ex. Megan Fox is less hot to me because I have heard she is a fucking cunt.	Highly toxic
Fuck all of you Jersey haters. I harbor no hate for any states, except those in the south...	
Fuck that. I love seeing bitches lose their shit and go full retard on each other.	
Fuck I hate people that delete or not have a facebook account...	Slightly toxic
She always does this hideous (like her face) fake laugh and says that I'm such a dumb blonde...	
...Other people do well so they must be evil rather than hard working smart sons-abitches, huh?	
I have pretty extreme ADHD and I do it to stay focused...	Nontoxic
I feel you on this one...with the exception of hard boiled egg in ramen...	
What I am saying is that the word "talent" is not that well defined...	

toxic (2.78% highly toxic and 11.04% slightly toxic). Some of the labeled comments from our collection are in Table 2. Due to space limitations, we only show relevant portions of the comments in the examples from Table 2.

Build models

An integral part of our methodology involves building a robust prediction model that can classify posts and comments based on the crowdsourced labels into three classes, which are (a) nontoxic, (b) slightly toxic, and (c) highly toxic. The following subsections describe the features extracted from the labeled dataset and the classification models we evaluated to predict toxicity at different levels.

Extracting features

For the prediction task, we utilized various features that characterize the semantic properties of text. First, we examined n-gram features at different configurations; then, we extracted an advanced set of features based on word embeddings, followed by a set of NLP-based features derived from the comments text. The following subsections explain each of the feature categories in more detail.

N-gram features. Before computing all the features, we cleaned the collection by removing new lines, URLs, and Reddit-specific (https://github.com/LoLei/redditcleaner; retrieved on Oct. 16, 2021) symbols like bullet points, quotes, strikethrough, spoilers, and coding parts from Reddit text. Furthermore, we split the collection text into tokens based on spaces and normalized all capital letters to lowercase letters. Then, we extracted 3,000 feature vectors from multiple variations of n-gram representations, including unigram features, bigram features, TF-IDF features, and n-gram features with a token range from three to five (*Yin et al., 2009*).

Word embedding features. For embedding features, we created vectors in a low-dimensional space using a distributed memory model from the vector representations of the cleaned tokens in the collection (*Le & Mikolov, 2014*). So, we used Python's Gensim library (https://radimrehurek.com/gensim/; retrieved on May 20, 2018) to build a skip-gram model and train the embeddings using a window size of 10, hierarchical softmax, and negative sampling of 10 noisy words. Then, we used the model to generate 300 word2vec feature vectors. Lastly, we trained another skip-gram model of window size 15, negative sampling of seven noisy words, and a learning rate of 0.03 to represent sentences as 300 doc2vec features.

NLP-based features. In addition to the previous features, we computed 37 shallow features based on natural language processing (NLP) techniques. Table 3 shows a summary of the list of features divided by the type of calculations we performed to obtain such features. We found that all the NLP-based features typically involve counting tokens or measuring the ratios of counts. Some of the features in Table 3 are adopted from *Salminen et al. (2018)*, where they used similar features to identify hateful posts on Facebook and YouTube.

Classify content

Classification based on classical machine learning

The classification approach considered several issues that persisted in the collection, such as the skewness of the classes. Since the labeled collection is highly skewed (86.81% of the comments are non-toxic), we had to address the class imbalance issue. One way to address this issue is to apply the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek links (an under-sampling technique) (*Batista, Prati & Monard, 2004*). SMOTE performs over-sampling, and links are cleaned by under-sampling using Tomek links. The next step in feature-based classification was to apply feature transformation by following a simple min-max scaling approach to normalize all features. Then, we performed feature selection to reduce the dimensionality of large feature vectors like n-grams. The last step in the classification procedure involved performing a grid search for parameter tuning followed by repeated stratified cross-validation over five folds.

Classification based on neural networks

Recurrent neural networks (RNNs) are artificial neural networks designed to process a sequence of inputs with temporal characteristics. One famous example is the long short-term memory (LSTM) neural network consisting of an input gate, a cell, a forget

Table 3 The list of 37 NLP-based features split into two categories based on the type of computation.

Feature types	List of features
Counts 20 feature	Characters (text length), words, capitals, nouns, verbs, adjectives, stop words, punctuations, periods, quotes, unknown words, discourse connectives, politeness words, rudeness words, single tokens, repeated punctuations, unique words, profane words, modal words, non alpha-numeric characters
Ratios (a:b) 17 features	a = counts of words, capitals, stop words, unique words, punctuations, nouns, verbs, adjectives b = text length
	a = counts of capitals, characters (without spaces), stop words, unique words, punctuations, profane words, nouns, verbs, adjectives b = count of words

gate, and an output gate. Another neural network type is the convolution neural network (CNN or ConvNet), which is commonly used to analyze images. However, a CNN can also be employed in other applications to detect data patterns (*Johnson & Zhang, 2016*).

In natural language processing, one of the most prominent deep learning networks is the transformer (*Devlin et al., 2019*), which handles sequences of ordered data to solve problems like machine translation. The difference between RNNs and transformers is that the latter do not require data to be processed in order. With this advantage, transformer models can be faster than RNNs during the training phase due to the parallelization of the data processing phase (*Vaswani et al., 2017*). A well-known transformer model is BERT, which consists of a multilayer bidirectional transformer encoder (*Devlin et al., 2019*).

Explore results

Toxicity judgments of user content

To determine user toxicity, we compute the percentages of highly toxic content and slightly toxic content. Combining these provides a general judgment of a user's toxic behavior, regardless of the toxicity level. Furthermore, combining toxicity levels compensates for the skewness of the dataset by increasing the amount of data that represents what is considered toxic. For instance, if a user u creates three posts, one labeled highly toxic, one slightly toxic, and one nontoxic, this user u is 67% toxic. Then, we use percentiles to describe users based on the proportion of toxicity in their generated content. The quartile values, which include the 25th, 50th, 75th, and 100th percentiles, capture the distribution of users based on the toxicity proportions in their content. For instance, if the 25th percentile is 10, it means that 25% of the time, the toxicity proportions in users' posts are below 10.

Toxicity changes of users in subreddits

Prior studies of the behavior of online users found that temporal features can help characterize undesirable behavior like trolling (*Cheng, Danescu-Niculescu-Mizil & Leskovec, 2015*). Thus, investigating the temporal aspects of toxic behavior is an interesting problem from the perspective of *Mathew et al. (2019)*; *Kaakinen, Oksanen & Rsnen (2018)*, and *Cheng, Danescu-Niculescu-Mizil & Leskovec (2015)*. Toward this goal, we studied toxicity changes by computing the toxicity difference of user content across all subreddits. In Eq. (2), we illustrate how to calculate the change (Δ) of toxicity per user. To get the percentage of toxic content, we combined the counts of highly toxic and slightly toxic content that

users made within a particular subreddit. Then, we divided this figure by the total number of content items posted by these users within that same subreddit. Next, we computed the difference by subtracting the highest toxicity proportion from each user's lowest toxicity within a particular year. Finally, from Eq. (2), we computed the differences in users' content from posts and comments over the years.

$$\Delta_u = \max_s\left(\frac{N(c_{toxic})^s_u}{N(c_{total})^s_u}\right) - \min_s\left(\frac{N(c_{toxic})^s_u}{N(c_{total})^s_u}\right). \tag{2}$$

$N(c_{toxic})^s_u$ is the number of toxic content (comments or posts) by user u in subreddit s, and $N(c_{total})^s_u$ is total number of content (comments or posts) by user u in subreddit s.

For instance, if a subreddit s has a total of 20,000 posts, and a user u posted 1,000 slightly toxic and 800 highly toxic posts, we add both highly and slightly toxic posts to get a total of 1,800 toxic posts. Then, we divide the total number of toxic posts by the total number of posts in s to get a toxicity percentage of 0.09. With this procedure, we continue to get all the toxicity percentages of u to calculate Δ like Eq. (2). Obtaining the toxicity percentage of all users within subreddits in the posts and comments collections is necessary for subsequent analysis in our study.

Link analysis of user content

Another way to investigate user behavior is by looking at the links in their content. Since our collection is rich in metadata, we extracted the URL field that includes the link (if any) that accompanies a post. Then, we searched for all the URLs from comments text to build another version of our dataset that includes an ID, a time stamp, and URLs. We included the time stamp in this version of the dataset to conduct some statistical hypothesis tests, such as the Granger causality test (*Mukherjee & Jansen, 2017*) to identify relationships between URLs in user content and the toxicity of said content.

FINDINGS

Classification results

In the following subsections, we show the results of several experiments that build and evaluate machine learning models for detecting toxicity at different levels.

Classical classification models

For this part of the analysis, we computed features to build and tune four widely used classic machine learning models: Logistic Regression, Random Forest, Decision Tree, and XGBoost, using the 10,083 comments from the labeled collection. We chose these four algorithms for their extensive usage in previous research on hate detection (*Badjatiya et al., 2017*; *Salminen et al., 2018*).

To handle class imbalance, we used SMOTE and Tomek Links on the training portion of the dataset (0.80 of the labeled collection). Then, we transformed features by scaling all values to a range between zero and one. Additionally, we used the Random Forest algorithm to perform the classification, ranking, and selection of features (*Vens & Costa, 2011*). The results depicted in Table 4 show the precision, recall, F_1, AUC, and classification accuracy

Table 4 The classification performance of each feature category across four different classifiers.

	Logistic regression					Random forest					Decision tree					XGBoost				
Features	***P***	***R***	F_1	***AUC***	***ACC.***	***P***	***R***	F_1	***AUC***	***ACC.***	***P***	***R***	F_1	***AUC***	***ACC.***	***P***	***R***	F_1	***AUC***	***ACC.***
Unigram	0.649	**0.692**	0.668	0.878	85.1	0.653	0.443	0.479	0.882	84.1	0.618	0.606	0.611	0.763	85.2	0.715	0.653	**0.679**	0.913	**88.1**
Bigram	0.572	0.389	0.379	0.625	57.1	0.416	0.347	0.327	0.609	81.5	0.499	0.374	0.379	0.547	80.9	0.572	0.413	0.435	0.668	82.1
N-gram (3-5)	0.358	0.394	0.292	0.558	39.2	0.386	0.406	0.274	0.569	39.1	0.392	0.342	0.104	0.515	13.3	0.410	0.378	0.384	0.594	76.9
TFIDF	0.624	0.671	0.645	0.892	84.6	0.643	0.594	0.606	0.883	85.3	0.602	0.620	0.609	0.782	84.7	0.673	0.636	0.653	0.910	87.3
NLP	0.393	0.446	0.365	0.660	54.5	0.419	0.399	0.406	0.644	75.3	0.370	0.399	0.231	0.568	26.4	0.424	0.377	0.383	0.620	78.4
Word2vec	0.477	0.579	0.492	0.789	67.4	0.578	0.456	0.487	0.799	81.1	0.393	0.441	0.384	0.645	57.8	0.573	0.522	0.543	0.810	81.3
Doc2vec	0.498	0.593	0.523	0.828	73.8	0.570	0.531	0.548	0.822	81.9	0.426	0.474	0.434	0.652	66.4	0.561	0.551	0.556	0.815	81.3
All Features	0.610	0.641	0.624	0.893	84.1	0.662	0.509	0.552	0.884	84.9	0.592	0.578	0.584	0.735	84.6	**0.732**	0.636	0.671	**0.924**	87.8

Notes.
P, precision; R, recall.

of every classifier, where accuracy measures the number of correctly predicted data items divided by the total number of predicted items.

Findings from Table 4 show the best performing classification model is XGBoost, followed closely by Logistic Regression. As for the best features, the results show that models perform best on the unigram features. However, all the features combined through concatenation with XGBoost showed the highest precision and AUC scores at 0.732 and 0.924, respectively. On the other hand, with Logistic Regression, unigram features achieved a recall score of 0.692. As for the F_1 and accuracy, XGBoost achieved the highest scores of 0.679 and 88.1% on the unigram features. The grid search of XGBoost showed that the best learning rate is 0.3, and the best number of estimators is 300. Moreover, feature selection reduced the dimensionality of all the combined features from 12,637 to 1,406, where the top selected features belonged to unigram and word embedding feature categories. This outcome aligns with the prior work done by *Nobata et al. (2016)*, where their best-performing feature categories on all their datasets were the n-gram and distributional semantic features.

Neural network models

Despite the outstanding performance of classic machine learning models, studies found that some neural network architectures can outperform classical machine learning models, especially with capturing long-range dependencies from textual data in hate detection problems (*Badjatiya et al., 2017*). Therefore, we chose to experiment with varying configurations of BERT as a basis for our trials with CNNs, RNNs (LSTM and biLSTM), and transformer networks.

- **CNN:** This model used a convolution layer along with global max pooling and batch normalization layers to normalize the layer dimensions and speed up the performance of the model (*Zhou et al., 2016*). The network deployed a learning rate of 0.00002. The optimizer was adam, and the maximum sequence length for tokenizing the training set was 384. The embedding features were from a pretrained BERT-medium model (https://huggingface.co/google/bert_uncased_L-8_H-512_A-8; retrieved on Oct. 20, 2021)

Table 5 Performance of the neural network models in terms of the macro precision, recall, F_1, AUC, and accuracy scores.

	Evaluation metrics				
Neural network models	***Precision***	***Recall***	***Macro F1***	***AUC***	***Accuracy***
CNN	0.6600	0.6172	0.6092	0.9231	87.12%
BiLSTM	0.6222	0.6629	0.6216	0.9336	86.42%
CNN+LSTM	0.6645	0.6966	0.6724	0.9380	87.51%
LSTM+CNN	0.7431	0.7212	0.7261	0.9570	89.99%
fine-tuned BERT	**0.7930**	**0.8034**	**0.7952**	**0.9629**	**91.27%**

- **bidirectional LSTM:** The model used a bidirectional LSTM layer with embedding features from BERT-medium. Additionally, the model had average pooling layers with dense layers and was trained using the same learning rate and sequence size as the previous CNN model.
- **CNN+LSTM:** This model consisted of four channels with convolution layers, global max pooling, and batch normalization. In addition, the end of each channel has an LSTM layer. The final model consists of the combined channels with added drop-out layers. The same BERT-medium features were used in this model with the same configurations.
- **LSTM+CNN:** This model used a bidirectional LSTM layer followed by a series convolution, global max pooling, and batch normalization layers. Like the previous models, BERT-medium was used to obtain the feature vectors.
- **fine-tuned BERT:** This transformer model used the uncased (*i.e.,* lower case) base model of BERT, which consists of 12 layers (also called transformer blocks), 768 hidden layers, 12 attention heads, and 110 million parameters. To fine-tune the model, we used a custom focal loss (*Lin et al., 2017*) function with gamma =2 and alpha =7 to account for class imbalance. Additionally, we computed class weights from each class's distribution of data points and used them to improve the training performance. As for the learning rate, we set it to 0.00003 and used the Adam optimizer.

We evaluated the performance of the neural networks on the labeled training set, where 80% of the data was for training while the rest was for testing and validation purposes. Then, we used the dataset's testing portion to evaluate the models' performance by measuring macro precision, recall, F_1, AUC, and accuracy scores, as in Table 5. The findings show that, clearly, the fine-tuned BERT model outperforms all neural network models, as it achieves a precision score of 0.7930, recall of 0.8034, an F_1 score of 0.7952, AUC score of 0.9629, and accuracy score of 91.27%.

One drawback of using a BERT model is that training was relatively slow. However, this issue can be solved by adjusting the batch size configuration. Comparing the performance of neural network models in Table 5 with the classical classification models in Table 4, the results showed that the fine-tuned BERT model outperformed all of the other models. Even though the performance of the BERT model might look too good to believe, the neural-network-based model has shown its high performance in toxic-classification tasks. For example, Google Jigsaw hosted a toxicity detection challenge

Table 6 Prediction results of the toxicity levels of the posts and comments of users.

Class	Posts (%)	Comments (%)
Highly toxic	1,794,115 (2.05%)	133,588,229 (6.06%)
Slightly toxic	6,364,092 (7.28%)	254,531,824 (11.54%)
Nontoxic	79,218,705 (90.66%)	1,817,233,194 (82.39%)

to classify toxic comments from Wikipedia in 2018, and the first ranked team reported an AUC score of 0.9890 (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557; retrieved on Mar. 25, 2022).

In summary, our fine-tuned BERT model will be used to detect toxicity in the remainder of this study.

Transferability of models across subreddits

To use the models to predict the toxicity of all the Reddit comments, we must ensure that the model trained by the data from r/AskReddit is transferrable to data from other subreddits (*Fortuna, Soler-Company & Wanner, 2021*). Toward this goal, we obtained a random sample of 1,000 comments from the remaining 99 subreddits (besides r/AskReddit). Then, we used crowdsourcing to label the comments for their level of toxicity. Finally, we used the same labeling job we described earlier to obtain the ground truth labels. Comparing the results from the prediction model and the crowdsourcing workers showed that the agreement between them was 94.2%, meaning that the model trained on r/AskReddit can be generalizable to other subreddits.

Detecting and determining the toxicity of content and users

To answer RQ1 (*How can the toxicity levels of users' content and users across different communities be detected?*), we infer the toxicity of the entire posts and comments collection by using our fine-tuned BERT model. Post toxicity was detected by concatenating post titles and body sections (if they existed). As for comments, toxicity was predicted directly from the comment text. The results of running the prediction model on the entire collection of 87,376,912 posts and 2,205,581,786 comments are in Table 6. The results show that, collectively, 17.61% of the posts were toxic (*i.e.,* both highly toxic and slightly toxic) and that the remaining 82.39% were nontoxic.

After obtaining the toxicity levels of user posts, we applied our method to judge user toxicity and get the total number of users (and their toxicity percentages) in every quartile, as shown in Table 7. For users who leave posts, the table shows that, in the 25th percentile, 26.27% of users had toxicity proportions in the range (1%, 5%]. As for the 50th percentile, 25.56% of users had toxicity proportions that fell in the range (5%, 9%]. Subsequently, in the 75th percentile, the toxicity proportions for 24.91% of users were in the range (9%, 15%]. Additionally, Table 7 shows that, in the 100th percentile (*i.e.,* the maximum quartile), the toxicity proportions in 23.26% of users were in the range (15%, 100%]. Therefore, among the four quartiles, the 25th percentile had the largest number of users; the average toxicity was 3.40%.

Table 7 Judgment of users' toxicity based on their predominant behavior. The judgments include the total number of users, their percentage (%), and the toxicity range of their posts and comments.

Users judgment	Posts (%) - toxicity range	Comments (%) - toxicity range
25th percentile	124,056 (26.27%) - (1%, 5%]	234,899 (26.55%) - (1%, 11%]
50th percentile	120,705 (25.56%) - (5%, 9%]	209,133 (23.64%) - (11%, 16%]
75th percentile	117,651 (24.91%) - (9%, 15%]	220,799 (24.96%) - (16%, 22%]
100th percentile	109,821 (23.26%) - (15%, 100%]	219,891 (24.85%) - (22%, 100%]

In the comments collection, for users in the 25th percentile, 26.55% had toxicity proportions in the range (1%, 11%]. Concerning the 50th percentile, 23.64% of users had toxicity proportions with a range of (11%, 16%]. The findings show that, in the 75th percentile, 24.96% of users had toxicity proportions with a range of (16%, 22%]. Lastly, our results show that, in the 100th percentile, 24.85% of users had toxicity proportions in the range (22%, 100%]. Simply put, in the 25th percentile of users who leave comments, about 27% of the users have an average toxicity of 7.77%.

Changes in users' toxicity across communities

Since some users do not show consistent toxic (or nontoxic) behavior in their content, with RQ2 (*Does the toxicity of users' behavior change (a) across different communities or (b) within the same community?*), we examine the content-based changes in users' toxicity. Here, we check if a change (or a multitude of changes) in the toxicity of users occurs within the same subreddit or across different subreddits.

To examine toxicity changes in this study, we devised two different change conditions to look for in the users' collection. These conditions come from the two possible judgments of users' posting behavior based on their content's toxicity. Based on our methodology, we judge users based on their contributions as (1) toxic (slightly toxic and highly toxic) or (2) nontoxic. Based on these judgments, the conditions that we identified are as follows:

- **Condition 1:** Change in the toxicity of a user's contribution from nontoxic to toxic. (**NT** $\rightarrow$**T**)
- **Condition 2:** Change in the toxicity of a user's contribution from toxic to nontoxic. (**T** $\rightarrow$**NT**)

Additionally, this experiment checked whether the conditions were met within the same subreddit or across different subreddits. Given the criteria for investigating the change in toxicity, we examined the entire history of users' content in the posts and comments collections. First, we sorted content from oldest to newest based on time stamps. Then, we used the toxicity prediction labels that we obtained from our prediction model to check for the change conditions. For example, suppose the first (*i.e.,* oldest or earliest) post of a user is nontoxic and the subsequent (*i.e.,* following or newer) post is toxic. In that case, this user exhibits a change in toxicity due to condition 1. This experiment considers whether the change happened in the same subreddit or across multiple subreddits. Thus, we flagged every user based on the exhibited condition and identified the location of the change on Reddit. Subsequently, we used a majority voting mechanism to get the number of users

Table 8 Total number of users (and their %) that satisfy the conditions and their locations on Reddit in the posts and comments collections.

	Submission users			Comment users		
Conditions	**Total**	**Same subreddit**	**Multiple subreddits**	**Total**	**Same subreddit**	**Multiple subreddits**
1. NT →T	23,946 (5.11%)	3,668 (15.32%)	16,377 (68.39%)	10,500 (1.19%)	2,608 (24.84%)	6,412 (61.07%)
2. T →NT	10,123 (2.16%)	1,468 (14.50%)	4,631 (45.75%)	7,891 (0.89%)	1,987 (25.18%)	4,572 (57.94%)
1. NT →T & 2. T →NT	435,000 (92.74%)	15,869 (3.65%)	419,131 (96.35%)*	866,228 (97.92%)	48,070 (5.55%)	818,158 (94.45%)**

Notes.
NT, Nontoxic; T, Toxic.
50.80% of 419,131* and 68.63% of 818,158** show change in the same and multiple subreddits.

that exhibited changes due to each condition and their locations on Reddit. We performed majority voting by getting a list of all the toxicity changes and their locations for every user. If most of the total changes for a user fall under a specific condition or location, then a user change is due to this particular condition. Moreover, any user who shows a single change in toxicity due to any condition was removed to avoid any issues that might arise by posting a single toxic post. For instance, if a user had two posts that showed a change in any of the conditions, we did not include it in the study. With this approach, we found that users can show at least two changes due to the one or two conditions at different locations on Reddit.

The majority voting technique allowed us to count users based on the overall change in their content. This approach resulted in 177,307 (30.68%) posting users and 727,587 (81.67%) commenting users that show changes due to conditions 1 or 2. To further analyze these users, the results depicted in Table 8 show the distribution of users who satisfy conditions within the same subreddit or across multiple subreddits in both posts and comments.

In Table 8, we measured the percentage of users that satisfy each condition along with the combined conditions and came up with compelling observations from each collection. For example, starting with the posts collection, we found that users show the most change due to conditions 1 and 2, where 96.35% of these users show a change across multiple subreddits. Additionally, 5.11% of these users show a change due to condition 1, where 68.39% of them show changes across multiple subreddits; similarly, the majority of users that satisfy condition 2 also show change across different subreddits. Specifically, 45.75% of these users show changes over multiple communities, meaning that most of the changes in toxicity among users that post occur in different communities.

As for the comments collection, just like the posts collection, users show the most change due to conditions 1 and 2, where 1.19% of these users show a change within the same subreddit. Moreover, 0.89% of these users show change due to condition 2, where 57.94% of them show changes across different subreddits. In other words, most commenting users show changes in their toxicity across different subreddits. So, concerning RQ2, findings show that engaging with multiple communities can cause users to exhibit changes in their toxic posting behavior.

To further illustrate the changes in toxic behavior, we count the total number of changes per user without considering the majority voting technique we used in Table 8. Then,

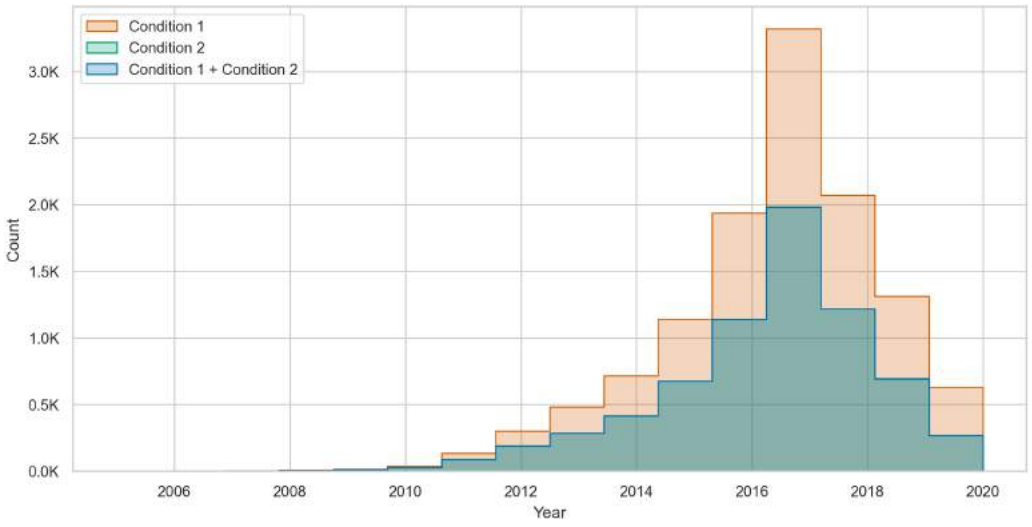


Figure 3 **The count of toxicity changes over time in posting users from condition 1 (NT → T), condition 2 (T → NT), and both conditions combined.**

Full-size DOI: 10.7717/peerjcs.1059/fig-3

we plot a histogram of the counts of changes that occur due to condition 1, condition 2, and both conditions combined. The histogram in Fig. 3 shows that over time, most of the changes are in condition 1, where posting users change their behavior from nontoxic to toxic.

Similarly, Fig. 4 shows that for commenting users, most of the changes occur due to condition 1. However, the gap between the counts of changes due to varying conditions is smaller in commenting users. Furthermore, we found that posting users can show up to 37,800 changes in toxicity while commenting users can show up to 295,912 changes due to both conditions. These high numbers of changes suggest that users change their toxicity when their volume of contributions increases. In fact, these values result from having at least two changes (*i.e.,* four posts or comments) from different conditions.

Changes in users' toxicity over time

Our collection's large volume of temporal data allows us to investigate toxicity changes over time. Therefore, we chose yearly intervals to answer RQ3 (*Does the toxicity of users change over time across different communities?*) and note observed changes in toxicity by computing the difference in the toxicity of every user across all the subreddits.

With Eq. (2), we computed users' toxicity percentages across all subreddits. Then, we calculated the change in toxicity (Δ) in every pair of years for posting and commenting users. Subsequently, we used scatter plots to visualize the change in toxicity per user across subreddits, which we then converted to heatmaps with varying smoothing parameters. The heatmap plots in Fig. 5 show the distribution of toxicity in the posts and comments with smoothing at 64 neighbors. Figures 5A and 5B show the heatmap plots for the posts from the years 2007–2008 and 2018–2019, respectively, while Figs. 5C and 5D show the heatmap plots for the comments collected from the years 2007–2008 and 2018–2019. To clarify the

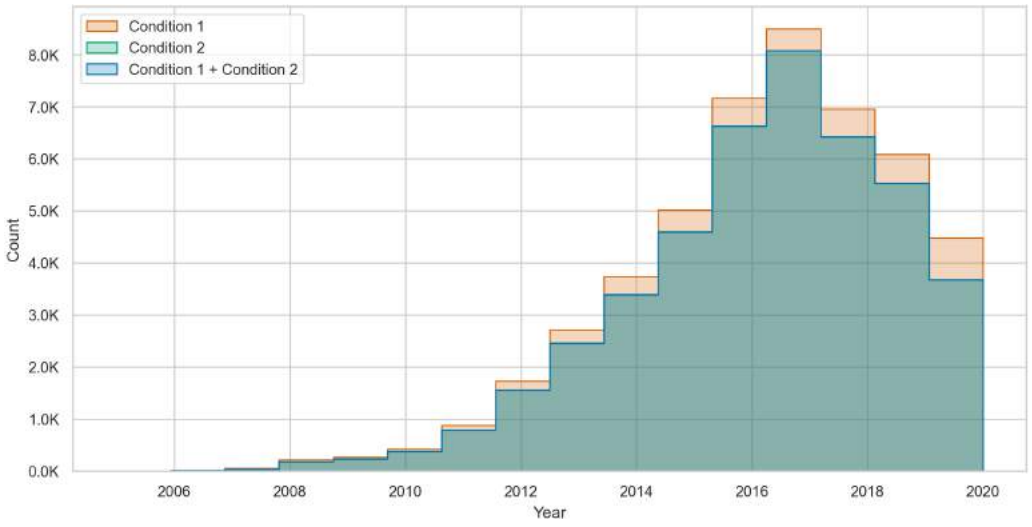


Figure 4 **The count of toxicity changes over time in commenting users from condition 1 (NT → T), condition 2 (T → NT), and both conditions combined.**

Full-size DOI: 10.7717/peerjcs.1059/fig-4

observations, we removed users who showed no change in their toxicity in the plots (*i.e.*, users with $\Delta = 0$). Due to space limitations, we only show the heatmap plots from two pairs of years representing the collection's beginning and ending periods. The overall temporal analysis of the content shows that over time, changes in the toxicity of users' posts disperse across the participating communities, as illustrated by the increase in dark color in Figs. 5B and 5D from Fig. 5.

In other words, over time, more users diffuse their toxic behavior to a large number of varying communities. To further support this finding with users who leave comments, we conducted a dependent T-test on the commenting users' deltas for 2007 and 2019 (*i.e.*, their initial delta and final delta). Results show that at $p < 0.001$, the t-statistic value of the users is 57.031, indicating a significant change in the posting behavior of users, which any of the conditions mentioned earlier can describe (change from toxic to non-toxic or change from non-toxic to toxic within the same subreddit).

Lastly, we visualize the Δ values of posting users in Fig. 6, where we show the total change per year and interpolate the change by computing the smoothed rolling average on intervals of three years. The average line shows that changes peak in 2017 but drop after this point, suggesting some form of stability in the behavior of posting users. Similarly, Fig. 7 shows that the average toxicity change in commenting users peaks in 2018 and drops slightly after this year. Unlike posting users, commenting users continue to show high amounts of change despite the drop after 2018.

Changes in toxicity and links

Originally, Reddit was a news-sharing website where users posted links to news websites or various multimedia content to instigate discussions with other users. Hence, most of Reddit's earlier content (primarily posts) contained links. However, links are not limited to

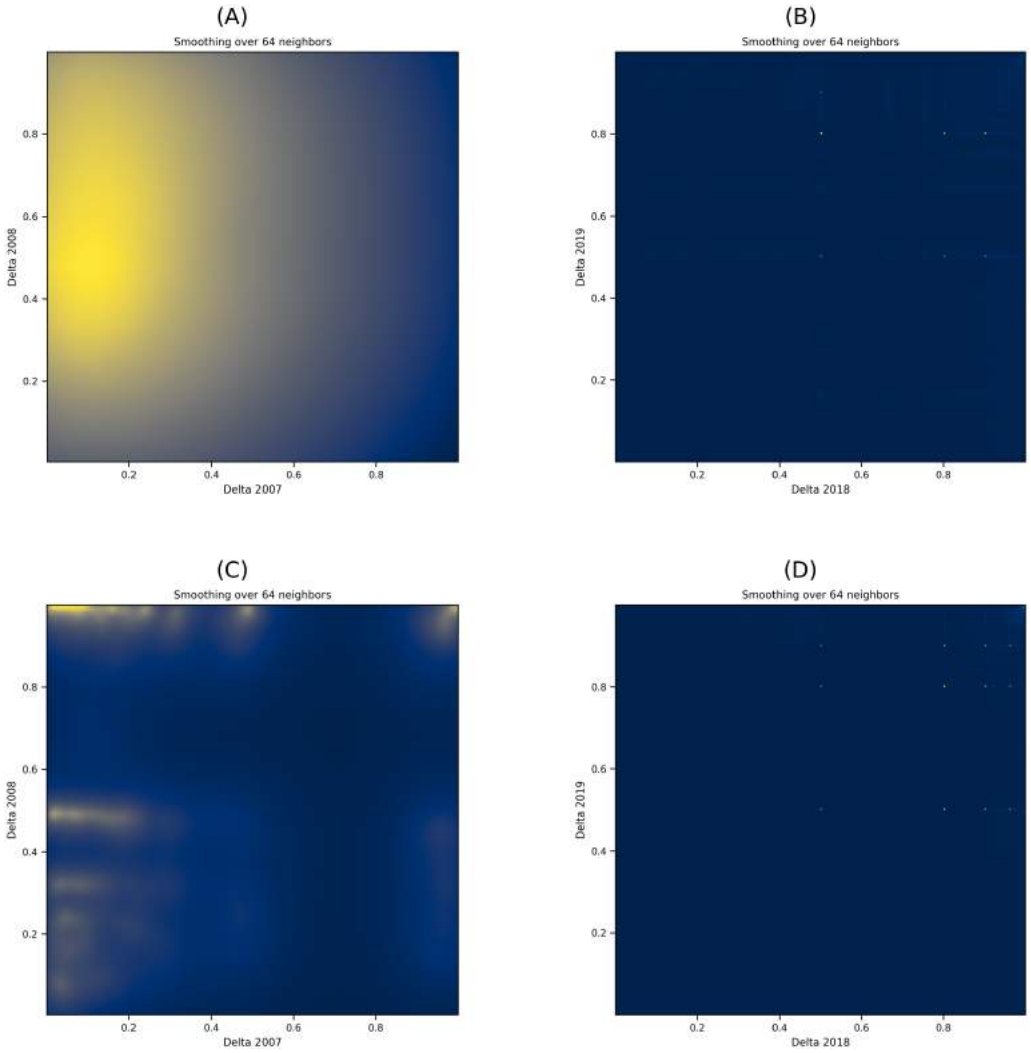


Figure 5 (A–D) Heatmap plots of the Δ in user posts and comments over two pairs of years. The dark color in the heatmap plot denotes scattered deltas while the light colors denotes concentrated deltas in specific locations.

Full-size DOI: 10.7717/peerjcs.1059/fig-5

posts, as users can also include different types of links in their comments. Since our earlier investigation of toxic behavior focused on the textual content of user content, it is only natural to examine links to identify any correlation between toxicity in user content and certain types of links. To begin investigating links, we performed a preliminary exploratory analysis on the entire collection to identify the number of links and the percentage of links from the total number of posts and comments per year. The statistics illustrated in the top portion of Fig. 8 show the total number of posts, toxic posts, and links in each year, while the bottom portion of the figure shows the corresponding normalized (*i.e.,* scaled) totals using the minimum and maximum values from the totals. The accompanying values from Fig. 8 are in Appendix C, where we also show the percentage of toxic posts in Table C1. Statistics from Table C1 show that between 2005 and 2012, more than 50% of user posts

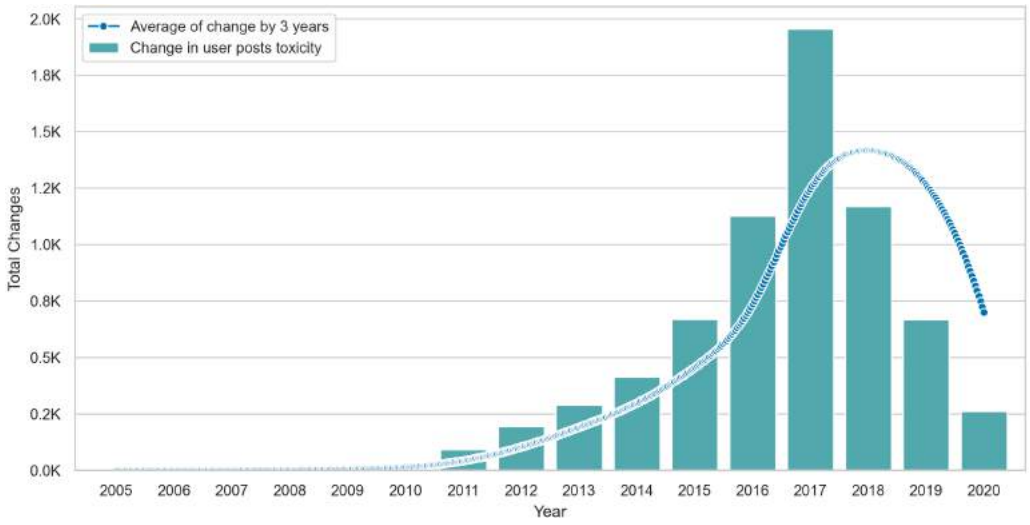


Figure 6 **The total amount of Δ in posting users content over time with an interpolation of Δ averages across three year intervals.**

Full-size DOI: 10.7717/peerjcs.1059/fig-6

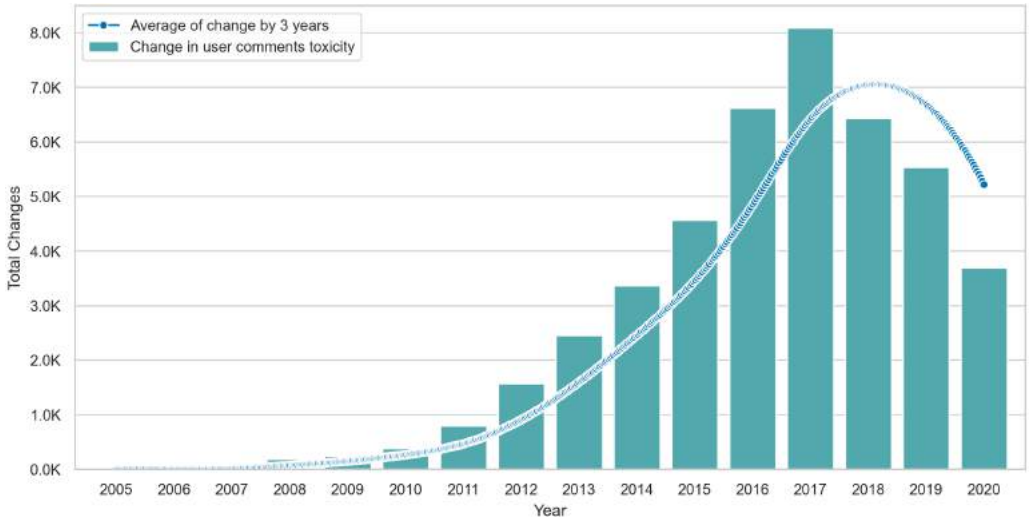


Figure 7 **The total amount of Δ in commenting users content over time with an interpolation of Δ averages across three year intervals.**

Full-size DOI: 10.7717/peerjcs.1059/fig-7

contained links, which means that, indeed, posts from the earlier years of Reddit contained a significant amount of links. Upon further investigation of the links in user content, we found external links that redirect to websites outside Reddit and internal links that redirect to a Reddit user, post, community, or multimedia content uploaded on Reddit servers. Additionally, we found that some of the links in posts refer to videos, images, or other types of identifiable media.

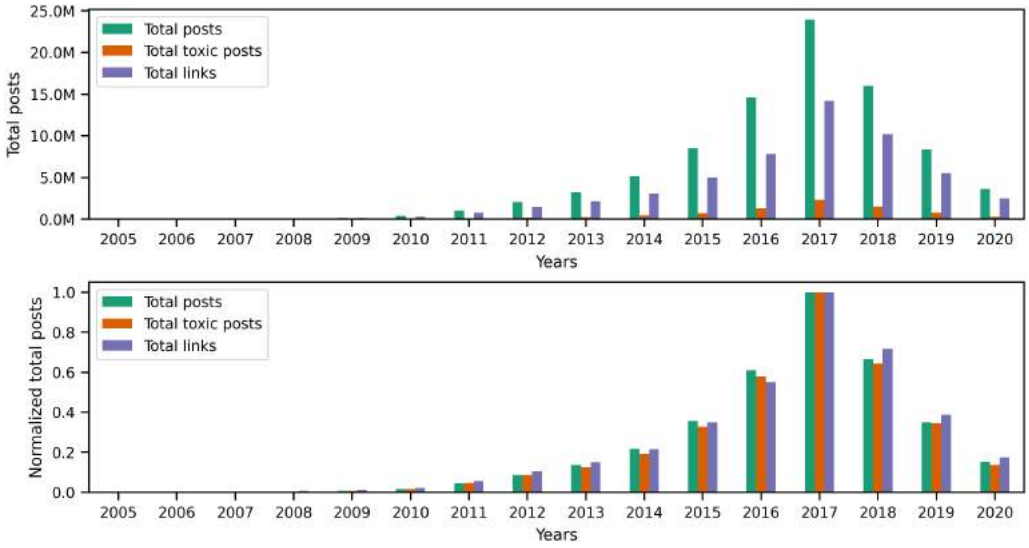


Figure 8 **The total number of posts, toxic posts, and links in every year followed by the normalized totals using the min-max scale.**

Full-size DOI: 10.7717/peerjcs.1059/fig-8

Our findings from the posts collection in Fig. 9 and Table C2 show that around 84% of the links in posts are external. As for the remaining internal links, we found that in posts, they typically link to an image uploaded to the i.redd.it domain, or a video uploaded to the v.redd.it domain, or a Reddit user. Figure 9 also shows the total number of links with identifiable media types. We used the mimetypes python module to guess the media types from the link's text representation (*i.e.,* path). So, if a link address ends with .mp4, mimetypes identify it as a video without examining the link's content. In posts, we found that the most identifiable media type happens to be images, so we calculated the total number of links with images and the percentage of images from the known media types in links. The results in Table C2 show that most of the links in posts from later years contain images.

As for the comments collection, the statistics in Fig. 10 and Table C3 show that the percentage of links in comments is significantly less than that of links in posts. However, this outcome does not diminish the fact that more than 110 million comments contain links, which is about 5% of the entire comments collection. Furthermore, just like the posts collection, Fig. 11 shows that comments from earlier years in the collection contain more links than comments in later years.

The results in Table C4 show that around 90% of the links in comments are external, and out of all the media types identifiable in these links, images seem to appear the most in user comments. However, when comparing the percentage of images in posts and comments, around 83% of links in posts contain images, while 72% of links in comments contain images. This observation makes sense because many communities on Reddit, such as r/cringepics, require users to post images in the community.

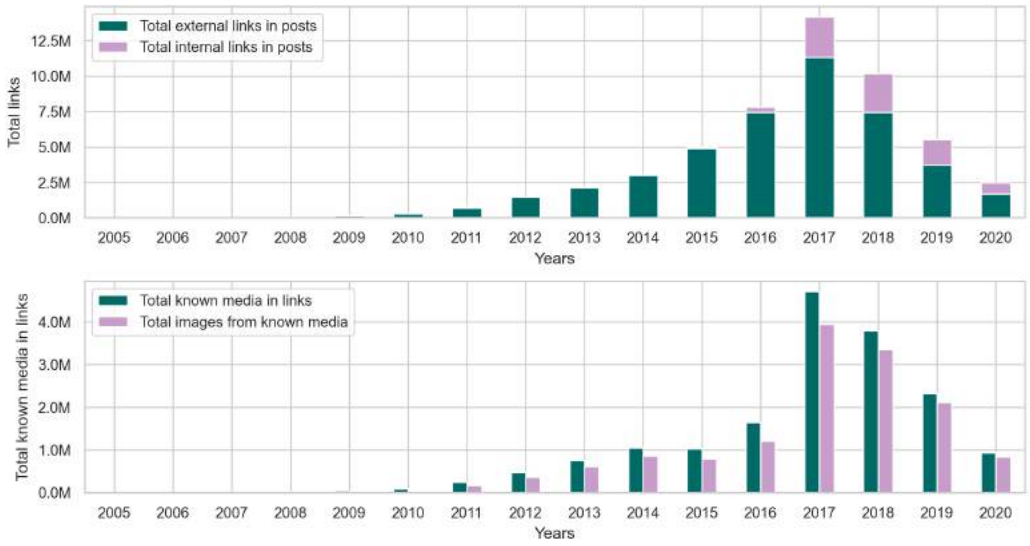


Figure 9 **The distribution of internal and external links, followed by the total number of known media types and image links from the posts collection.**

Full-size DOI: 10.7717/peerjcs.1059/fig-9

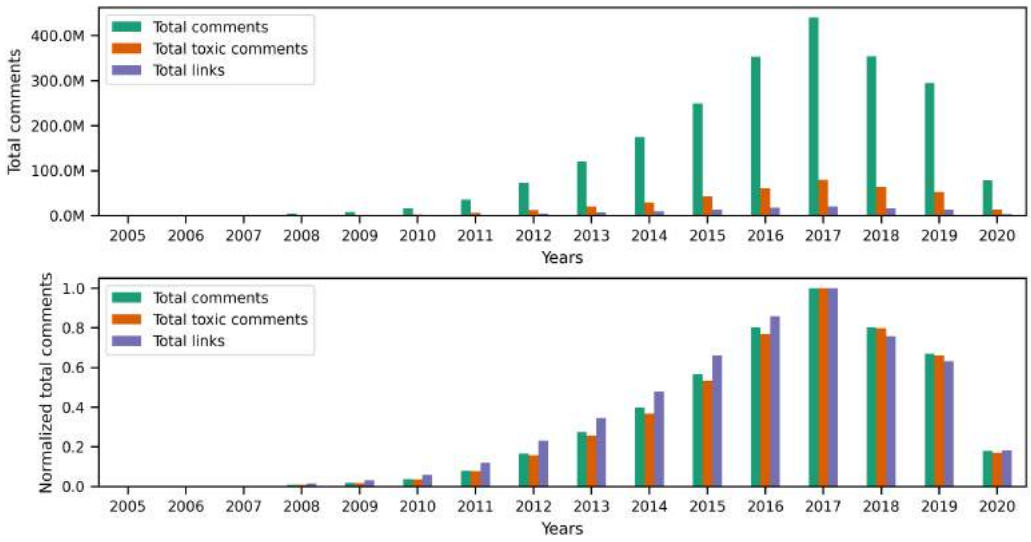


Figure 10 **The total number of comments, toxic comments, and links in every year followed by the normalized totals using the min-max scale.**

Full-size DOI: 10.7717/peerjcs.1059/fig-10

After performing the preliminary exploratory analysis of links in the collection, we used the Granger causality test to find correlations between toxic behavior and links in posts and comments. First, we conducted a test between the volume of content (X) in each collection and the volume of toxic (both highly and slightly toxic) content (Y). Then, we conducted another test between the volume of links in each collection (Z) and the volume of toxic (both highly and slightly toxic) content (Y). While our original intention was to perform

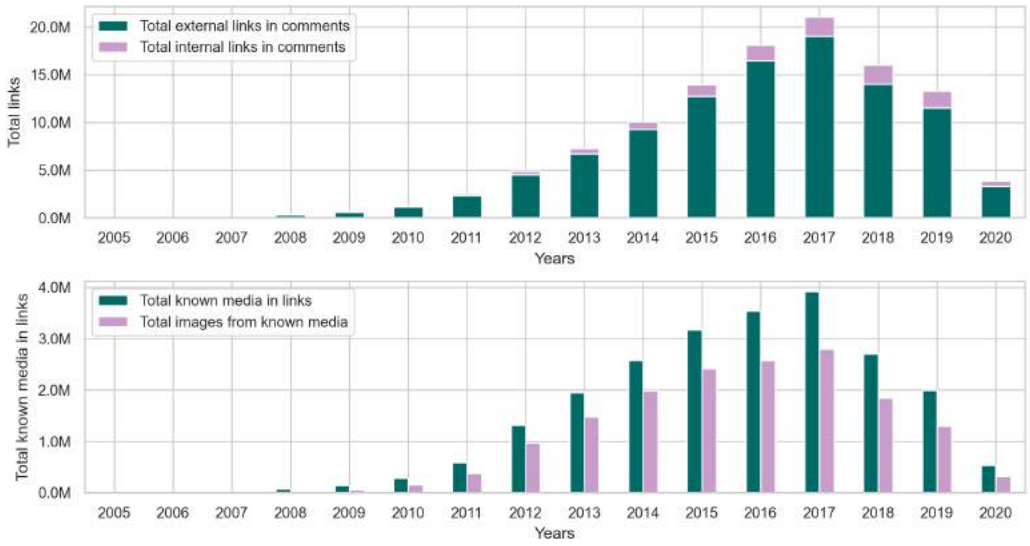


Figure 11 **The distribution of internal and external links, followed by the total number of known media types and image links from the comments collection.**

Full-size DOI: 10.7717/peerjcs.1059/fig-11

the Granger causality test on user posts as well, we found that, since almost 99% of posts contain links, running the test will not provide us with valuable insights on the relationship between links in posts and toxicity. Moreover, our yearly time series in user posts did not produce a stationary series, which does not satisfy the requirement to conduct the Granger causality test. Therefore, we limit our experiments to user comments. Table 9 shows the F-statistic, p-value, and selected minimum lags in years from conducting the Granger causality test on the volumes of comments, toxic comments, and links in comments. The causality of toxicity in user comments can be observed in Table 9, where the p-valu $e < 0.05$ for the volume of comments and links. In other words, the volume of comments and links in comments influences the volume of toxic comments in the collection.

IMPLICATIONS AND CONCLUSIONS

In this research, using over 10 thousand labeled comments, we trained feature-based and neural network-based models to infer the toxicity of user posts. We then used a fine-tuned BERT model to analyze a large-scale Reddit collection of more than two billion posts from more than 1.2 million users over 16 years for a detailed analysis of the toxic behavior of users in multiple online communities. Our contributions are three-fold:

- First, to our knowledge, we built one of the biggest labeled datasets of Reddit toxic comments and made it publicly available for further research (https://github.com/Hind-Almerekhi/toxicityChangesReddit). Additionally, compared to other binary labeled datasets, our dataset contains three levels of toxicity for each comment, ranging from non-toxic to slightly toxic to highly toxic.

Table 9 Results of Granger causality for the comments collection at a minimum lag in years.

	F-statistic	*p*-value	Lags (years)
(X,Y): X → Y	7.6306	0.014	2
(Z,Y): Z → Y	10.3849	0.014	3

- Second, by systematic comparisons of common feature-based models and neural network-based models, we demonstrate that a fine-tuned BERT model performs best for toxicity detection in posts and comments from Reddit.
- Third, our work is one of the first large-scale studies that investigate toxic behavior across multiple communities. We start with a list of cross-community users from the top 100 subreddits and expand our collection by obtaining posts and comments from more than 107,000 subreddits to reveal how users behave across communities from the perspective of toxicity.

Implications

Our work has several implications for the safety and moderation of online communities. These implications include the following:

Early detection of changes in toxicity

The dissemination of toxicity in online communities impacts the positive experience many users seek when using social media platforms. Several research studies showed that users could negatively influence each other when interacting in online communities (*Kwon & Gruzd, 2017*; *Zhang et al., 2018*; *Cheng et al., 2017*).

This type of negative behavior can continue to spread and harm online communities. Monitoring the change in users' toxicity can be an early detection method for toxicity in online communities. The proposed methodology can identify when users exhibit a change by calculating the toxicity percentage in posts and comments. This change, combined with the toxicity level our system detects in users' posts, can be used efficiently to stop toxicity dissemination. Furthermore, our methodology supports detecting toxicity early in online communities from users' toxicity. In an active setting, users' toxicity percentages can issue early alerts to online community moderators (bots or humans) so they can investigate potential toxicity incidents and take necessary actions to mitigate the further spread of toxicity in communities.

Aid moderators with toxicity changes

Judging the toxicity of user content may not always be ideal for preventing the spread of hate and incivility (*Rajadesingan, Resnick & Budak, 2020*). Our study showed that users changed their posts' toxicity within and across different communities. This change can result from fluctuations in the users' feelings or changes in the atmosphere of their communities. This change, coupled with the toxicity of the users' content, can create an accurate assessment method to prevent the spread of toxicity. For instance, instead of banning a user for a tasteless contribution they left once, moderators can consider the users' predominant toxicity and that of their previous content. This approach will

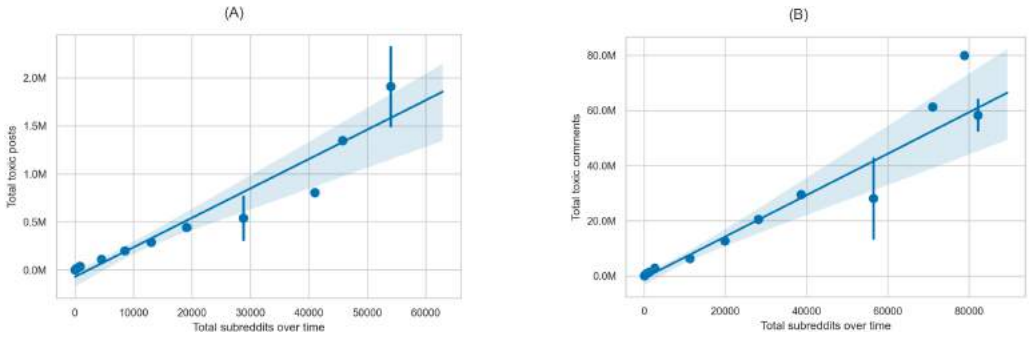


Figure 12 **Correlation between the total number of participating subreddits over time and (A) the total number of toxic posts and (B) the total number of toxic comments.**
Full-size DOI: 10.7717/peerjcs.1059/fig-12

prevent automated bots and moderators from excessively penalizing or banning users. This sophisticated user- and content-based toxicity assessment allows moderators to control toxicity and detect malicious users who deserve banning from online communities. Trolls and the like (*Cheng et al., 2017*) can also be prevented from polluting online communities by using our recommended method to judge users based on their content's predominant toxicity.

Moreover, the rules and norms of communities can be changed to prevent the spread of toxicity (*Chandrasekharan et al., 2018*). For example, when users show a rapid change in the toxicity of their behavior, the moderation ecosystem might raise alerts/reminders for any breaches that do not conform to the community norms. Lastly, our study suggests that one way to limit the spread of toxicity is by limiting the spaces (*i.e.,* communities) in which users can participate. To illustrate this finding, in Fig. 12 we show the correlation between the total amount of toxic posts (Fig. 12A), comments (Fig. 12B) and the total number of communities that users participate in over time. The figures show a positive correlation between the increase in the number of communities and the increase in toxicity. Ultimately, we cannot guarantee that this is the only reason behind the increase in toxic content, yet we argue that increasing communities could allow users to spread toxic content.

Limitations and future work

Since our research focuses mainly on text analysis to detect toxicity, one limitation is that toxicity takes different forms (*e.g.*, images, videos, and sound clips). While more sophisticated techniques are required to examine and analyze such content (*Rafiq et al., 2015*), multimedia submissions also have text titles that we studied in this work. Another limitation of this work is that it does not fully consider bias in the toxicity of the labels we obtained through crowdsourcing (*Vaidya, Mai & Ning, 2020*; *Hanu, Thewlis & Haco, 2021*). However, since toxicity is a subjective matter, our study performed toxicity detection in a simplified manner without accounting for subjectivity (*Zhao, Zhang & Hopfgartner, 2022*). Lastly, we note that our study did not tackle any contextual or categorical characteristics of toxic content (*Radfar, Shivaram & Culotta, 2020*). That is partially due to the heterogeneous nature of most Reddit communities, making it extremely

difficult to capture their context to judge different types of content, such as profanity in certain Not Safe For Work (NSFW) communities (*Madukwe & Gao, 2019*).

Upon investigating the toxic posting behavior of users, we came across several ideas that can lead to interesting future research directions. One of the ideas focuses on different scenarios involving users joining new communities and considering the changes in their toxicity to these new communities (*Cheng, Danescu-Niculescu-Mizil & Leskovec, 2015*; *Choi et al., 2015*; *Cheng et al., 2017*). Another take on the problem of toxic posting behavior can focus on specific topics within each community (*e.g.*, controversial topics or hot news) to study how they trigger toxicity within users, as opposed to noncontroversial or regular topics (*e.g.*, entertainment news or funny stories). Besides focusing on various topics within online communities, one can also study the temporal characteristics that foster the evolution of toxic communities from a few users with predominantly toxic posts. Considering the factors that contribute to the rapid growth of such toxic communities is also necessary for providing moderators and platform designers with the right tools to prevent toxicity from contaminating online communities.

CONCLUSIONS

In this research, we investigated users' toxic cross-community behavior based on the toxicity of their posts and comments. Our fine-tuned BERT model achieved a classification accuracy of 91.27% and an average F_1 score of 0.79, showing a 2% and 7% improvement in performance compared to the best-performing baseline models based on neural networks. We addressed RQ1 by running a prediction experiment on the posts and comments from our Reddit collection. The analysis showed that 9.33% of the posts are toxic, and 17.6% of the comments are toxic. We answered RQ2 by investigating the changes in the toxicity of users' content across communities based on two primary conditions. First, our analysis showed that 30.68% of posting users showed changes in their toxicity levels. Moreover, 81.67% of commenting users showed changes in their toxicity levels, mainly across multiple communities. Moreover, we found through answering RQ3 that, over time, toxicity disperses with an increase in the number of participating users and the frequency of cross-community participation. This finding is helpful because it can provide community moderators with leads to help them track patterns from active users to prevent them from spreading toxic content online.

Lastly, we conducted a Granger causality test between the volume of comments, the volume of links in comments, and the volume of toxicity. We found that links in comments can influence toxicity within those comments. This research addresses a prominent issue in social media platforms: toxic behavior negatively impacts other users' experience. Thus, we believe it is necessary to conduct more research on users' toxic behavior to help us understand the behavior's dynamics.

APPENDIX A

Table A1 The top 100 (1–50) subreddits ranked by the total number of subscribers along with the total number of posts and comments from our dataset.

Rank	Subreddit	Subscribers	Posts	Comments
1	r/funny	17,934,343	1,226,923	38,340,865
2	r/AskReddit	17,829,339	3,571,863	176,219,659
3	r/todayilearned	17,658,854	379,675	24,982,380
4	r/science	17,599,931	133,623	4,387,503
5	r/worldnews	17,522,963	613,496	31,844,773
6	r/pics	17,517,767	905,005	35,795,690
7	r/IAmA	17,215,966	34,839	7,650,945
8	r/gaming	16,860,176	762,776	25,502,647
9	r/videos	16,758,252	838,316	23,327,414
10	r/movies	16,468,377	407,215	17,824,942
11	r/aww	15,907,509	629,697	8,902,569
12	r/Music	15,879,988	400,517	4,567,182
13	r/gifs	15,051,932	175,824	12,300,726
14	r/news	14,994,220	896,236	25,720,157
15	r/explainlikeimfive	14,671,688	343,303	6,037,936
16	r/askscience	14,587,860	253,372	1,609,707
17	r/EarthPorn	14,197,666	92,811	1,178,788
18	r/books	13,699,914	83,049	2,786,204
19	r/television	13,617,822	129,974	6,005,034
20	r/LifeProTips	13,212,746	113,033	3,541,704
21	r/mildlyinteresting	13,170,674	319,910	7,369,887
22	r/space	12,915,830	77,426	2,126,598
23	r/Showerthoughts	12,797,856	944,483	9,315,365
24	r/DIY	12,753,699	52,804	1,647,926
25	r/Jokes	12,604,471	290,732	2,985,167
26	r/sports	12,552,016	129,084	2,422,532
27	r/gadgets	12,518,386	43,995	1,420,475
28	r/tifu	12,504,386	53,468	3,271,900
29	r/nottheonion	12,451,389	112,276	4,024,654
30	r/InternetIsBeautiful	12,433,228	29,919	338,448
31	r/photoshopbattles	12,363,401	93,001	624,593
32	r/history	12,356,415	48,904	1,105,441
33	r/food	12,351,889	181,129	2,401,415
34	r/Futurology	12,332,702	88,915	3,437,276
35	r/Documentaries	12,298,293	47,042	1,539,521
36	r/dataisbeautiful	12,293,355	36,084	2,016,558
37	r/listentothis	12,243,720	143,301	379,788
38	r/UpliftingNews	12,213,060	56,064	1,495,984
39	r/personalfinance	12,212,893	139,922	4,497,818

(*continued on next page*)

Table A1 (*continued*)

Rank	Subreddit	Subscribers	Posts	Comments
40	r/GetMotivated	12,135,696	50,356	935,431
41	r/OldSchoolCool	12,083,017	107,293	3,061,358
42	r/philosophy	12,081,626	23,948	713,078
43	r/Art	11,868,112	151,648	926,225
44	r/nosleep	11,678,653	23,966	607,853
45	r/creepy	11,665,125	46,469	1,147,867
46	r/WritingPrompts	11,612,067	206,881	948,065
47	r/TwoXChromosomes	11,215,698	52,283	3,002,720
48	r/Fitness	6,186,196	115,007	4,672,605
49	r/technology	5,551,587	267,264	9,225,651
50	r/WTF	4,861,274	239,730	19,032,277

Table A2 The top 100 (51–100) subreddits ranked by the total number of subscribers along with the total number of posts and comments from our dataset.

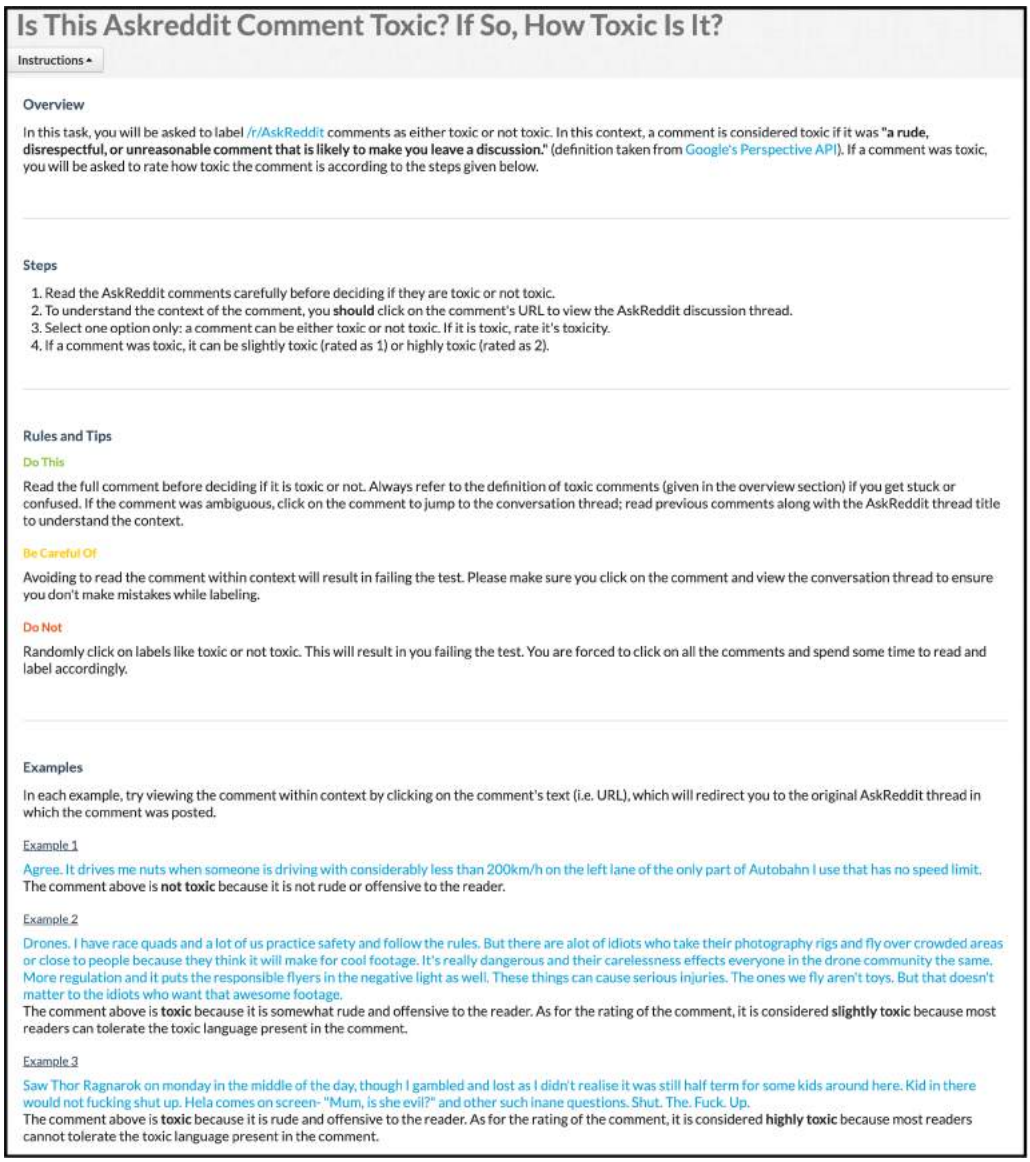
Rank	Subreddit	Subscribers	Submissions	Comments
51	r/bestof	4,772,718	41,570	1,869,232
52	r/AdviceAnimals	4,322,195	491,179	17,734,082
53	r/politics	3,468,561	955,613	60,494,833
54	r/atheism	2,096,408	147,972	8,890,785
55	r/europe	1,526,462	123,920	5,707,891
56	r/interestingasfuck	1,385,740	73,778	3,595,720
57	r/woahdude	1,345,016	58,643	1,514,561
58	r/leagueoflegends	1,118,408	550,318	22,450,314
59	r/gameofthrones	1,116,208	110,096	4,023,561
60	r/pcmasterrace	1,103,955	392,241	11,081,807
61	r/BlackPeopleTwitter	1,073,938	67,891	5,072,940
62	r/reactiongifs	1,038,629	86,617	1,488,555
63	r/trees	1,006,481	287,696	6,251,863
64	r/Unexpected	965,760	45,269	1,404,524
65	r/Overwatch	948,162	329,076	6,781,358
66	r/oddlysatisfying	905,675	53,992	1,683,927
67	r/Android	897,620	113,772	6,668,499
68	r/wholesomememes	840,077	36,283	1,003,541
69	r/Games	839,529	158,645	8,505,029
70	r/programming	826,809	51,209	3,208,560
71	r/4chan	819,656	38,445	2,338,530
72	r/nba	805,171	295,232	28,721,576
73	r/facepalm	791,286	36,112	1,955,898
74	r/cringepics	780,791	30,813	2,132,087
75	r/me_irl	779,311	435,999	2,081,621
76	r/relationships	774,812	61,107	5,975,894
77	r/sex	761,247	39,179	2,305,426
78	r/pokemon	760,949	120,266	3,992,989
79	r/fffffffuuuuuuuuuuuu	759,747	53,360	2,372,289
80	r/lifehacks	755,376	9,845	480,229
81	r/Frugal	741,976	24,157	1,611,244
82	r/soccer	736,005	231,161	20,685,778
83	r/tattoos	732,943	29,924	470,543
84	r/pokemongo	730,140	115,926	2,434,360
85	r/comics	726,976	85,320	1,128,707
86	r/OutOfTheLoop	688,156	55,979	1,249,050
87	r/malefashionadvice	684,010	56,821	2,314,703
88	r/CrappyDesign	667,846	78,494	1,376,476
89	r/StarWars	658,622	121,300	3,685,999
90	r/YouShouldKnow	644,359	8,068	577,679

(*continued on next page*)

Table A2 (*continued*)

Rank	Subreddit	Subscribers	Submissions	Comments
91	r/AskHistorians	637,383	106,763	629,207
92	r/buildapc	635,055	281,713	5,049,608
93	r/nfl	626,637	197,693	33,747,257
94	r/HistoryPorn	626,507	33,838	647,963
95	r/RoastMe	622,922	23,436	1,808,683
96	r/loseit	613,079	47,978	1,388,423
97	r/FoodPorn	612,361	39,742	570,383
98	r/AnimalsBeingJerks	605,103	11,596	429,629
99	r/dankmemes	598,376	238,182	2,003,893
100	r/rickandmorty	586,805	51,538	1,017,558

APPENDIX B



Is This Askreddit Comment Toxic? If So, How Toxic Is It?

Instructions ▲

Overview

In this task, you will be asked to label /r/AskReddit comments as either toxic or not toxic. In this context, a comment is considered toxic if it was **"a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."** (definition taken from Google's Perspective API). If a comment was toxic, you will be asked to rate how toxic the comment is according to the steps given below.

Steps

1. Read the AskReddit comments carefully before deciding if they are toxic or not toxic.
2. To understand the context of the comment, you **should** click on the comment's URL to view the AskReddit discussion thread.
3. Select one option only: a comment can be either toxic or not toxic. If it is toxic, rate it's toxicity.
4. If a comment was toxic, it can be slightly toxic (rated as 1) or highly toxic (rated as 2).

Rules and Tips

Do This

Read the full comment before deciding if it is toxic or not. Always refer to the definition of toxic comments (given in the overview section) if you get stuck or confused. If the comment was ambiguous, click on the comment to jump to the conversation thread; read previous comments along with the AskReddit thread title to understand the context.

Be Careful Of

Avoiding to read the comment within context will result in failing the test. Please make sure you click on the comment and view the conversation thread to ensure you don't make mistakes while labeling.

Do Not

Randomly click on labels like toxic or not toxic. This will result in you failing the test. You are forced to click on all the comments and spend some time to read and label accordingly.

Examples

In each example, try viewing the comment within context by clicking on the comment's text (i.e. URL), which will redirect you to the original AskReddit thread in which the comment was posted.

Example 1

Agree. It drives me nuts when someone is driving with considerably less than 200km/h on the left lane of the only part of Autobahn I use that has no speed limit.
The comment above is **not toxic** because it is not rude or offensive to the reader.

Example 2

Drones. I have race quads and a lot of us practice safety and follow the rules. But there are alot of idiots who take their photography rigs and fly over crowded areas or close to people because they think it will make for cool footage. It's really dangerous and their carelessness effects everyone in the drone community the same. More regulation and it puts the responsible flyers in the negative light as well. These things can cause serious injuries. The ones we fly aren't toys. But that doesn't matter to the idiots who want that awesome footage.
The comment above is **toxic** because it is somewhat rude and offensive to the reader. As for the rating of the comment, it is considered **slightly toxic** because most readers can tolerate the toxic language present in the comment.

Example 3

Saw Thor Ragnarok on monday in the middle of the day, though I gambled and lost as I didn't realise it was still half term for some kids around here. Kid in there would not fucking shut up. Hela comes on screen- "Mum, is she evil?" and other such inane questions. Shut. The. Fuck. Up.
The comment above is **toxic** because it is rude and offensive to the reader. As for the rating of the comment, it is considered **highly toxic** because most readers cannot tolerate the toxic language present in the comment.

Figure B1 **The labeling task instructions that we provided to crowd workers.**
Full-size DOI: 10.7717/peerjcs.1059/fig-B1

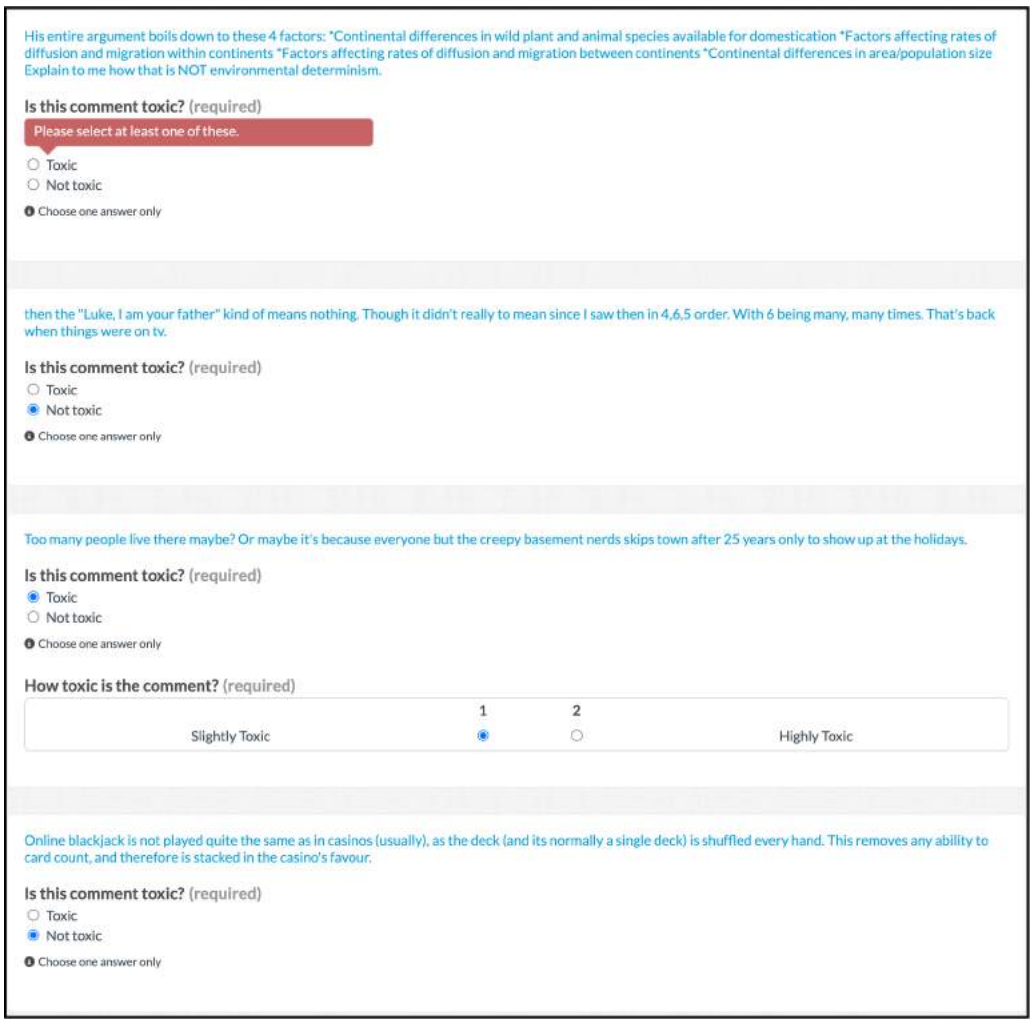


Figure B2 **The validation questions that crowd workers had to pass before beginning labeling.**
Full-size DOI: 10.7717/peerjcs.1059/fig-B2

APPENDIX C

Table C1 The total number of posts, toxic posts, and links in every year followed by the normalized totals using the min-max scale.

	Original counts			Normalized counts		
Year	Posts	Toxic (%)	Links (%)	Posts	Toxic	Links
2005	1,690	81 (4.79)	1,690 (100)	0.0	0.0	0.0
2006	10,917	895 (8.2)	10,917 (100)	0.0004	0.0004	0.0006
2007	47,556	4,372 (9.19)	47,556 (100)	0.0019	0.0018	0.0032
2008	105,825	10,407 (9.83)	101,357 (95.78)	0.0043	0.0044	0.007
2009	188,171	18,344 (9.75)	167,255 (88.88)	0.0078	0.0079	0.0117
2010	385,611	37,612 (9.75)	305,616 (79.26)	0.016	0.0162	0.0214
2011	1,058,576	108,253 (10.23)	791,010 (74.72)	0.0441	0.0466	0.0556
2012	2,052,406	201,278 (9.81)	1,472,632 (71.75)	0.0856	0.0866	0.1036
2013	3,247,906	287,250 (8.84)	2,133,632 (65.69)	0.1355	0.1236	0.1501
2014	5,176,179	442,283 (8.54)	3,054,626 (59.01)	0.2159	0.1904	0.215
2015	8,532,341	762,746 (8.94)	4,978,694 (58.35)	0.3559	0.3284	0.3505
2016	14,613,378	1,346,608 (9.21)	7,822,994 (53.53)	0.6097	0.5797	0.5508
2017	23,967,825	2,322,698 (9.69)	14,201,243 (59.25)	1.0	1.0	1.0
2018	15,968,062	1,496,033 (9.37)	10,201,097 (63.88)	0.6662	0.6441	0.7183
2019	8,379,712	804,135 (9.6)	5,523,004 (65.91)	0.3496	0.3462	0.3888
2020	3,640,757	315,212 (8.66)	2,468,515 (67.8)	0.1518	0.1357	0.1737

Table C2 The total number of internal links, external links, known media type links, and image links from the posts collection.

	Categories of links		Contents of links	
Year	Internal	External	Known media (%)	Images (%)
2005	0	1,690	565 (33.43)	9 (1.59)
2006	0	10,917	3,392 (31.07)	149 (4.39)
2007	0	47,556	13,001 (27.34)	1,267 (9.75)
2008	57	101,300	28,890 (28.5)	3,272 (11.33)
2009	197	167,058	45,240 (27.05)	8,270 (18.28)
2010	282	305,334	93,788 (30.69)	38,803 (41.37)
2011	81,426	709,584	245,682 (31.06)	166,934 (67.95)
2012	550	1,472,082	471,553 (32.02)	364,726 (77.35)
2013	1,078	2,132,554	751,834 (35.24)	610,537 (81.21)
2014	30,615	3,024,011	1,052,430 (34.45)	856,061 (81.34)
2015	70,917	4,907,777	1,032,840 (20.75)	788,774 (76.37)
2016	389,853	7,433,141	1,643,321 (21.01)	1,211,723 (73.74)
2017	2,862,721	11,338,522	4,710,886 (33.17)	3,940,218 (83.64)
2018	2,757,591	7,443,506	3,788,301 (37.14)	3,354,431 (88.55)
2019	1,803,981	3,719,023	2,323,104 (42.06)	2,109,346 (90.8)
2020	771,499	1,697,016	934,850 (37.87)	842,602 (90.13)

Table C3 The total number of comments, toxic comments, and links in every year followed by the normalized totals using the min-max scale.

	Original counts			Normalized counts		
Year	**Comments**	**Toxic (%)**	**Links (%)**	**Comments**	**Toxic**	**Links**
2005	310	26 (8.39)	38 (12.26)	0.0	0.0	0.0
2006	169,608	17,553 (10.35)	19,052 (11.23)	0.0004	0.0002	0.0009
2007	849,828	119,477 (14.06)	74,055 (8.71)	0.0019	0.0015	0.0035
2008	4,573,561	795,377 (17.39)	338,229 (7.4)	0.0104	0.01	0.0161
2009	8,494,022	1,450,220 (17.07)	674,146 (7.94)	0.0193	0.0181	0.032
2010	16,384,988	2,830,121 (17.27)	1,269,001 (7.74)	0.0372	0.0354	0.0602
2011	35,473,547	6,260,102 (17.65)	2,531,916 (7.14)	0.0806	0.0783	0.1202
2012	72,943,244	12,704,594 (17.42)	4,883,585 (6.7)	0.1657	0.1589	0.2318
2013	121,155,630	20,530,863 (16.95)	7,262,543 (5.99)	0.2752	0.2569	0.3447
2014	175,223,888	29,461,807 (16.81)	10,063,833 (5.74)	0.398	0.3686	0.4777
2015	249,496,457	42,597,414 (17.07)	13,934,125 (5.58)	0.5667	0.5329	0.6614
2016	352,996,950	61,346,847 (17.38)	18,100,953 (5.13)	0.8017	0.7675	0.8591
2017	440,297,137	79,930,465 (18.15)	21,068,587 (4.79)	1.0	1.0	1.0
2018	353,701,991	63,881,250 (18.06)	15,995,772 (4.52)	0.8033	0.7992	0.7592
2019	294,450,367	52,818,942 (17.94)	13,309,224 (4.52)	0.6688	0.6608	0.6317
2020	79,370,258	13,603,534 (17.14)	3,867,698 (4.87)	0.1803	0.1702	0.1836

Table C4 The total number of internal links, external links, known media type links, and image links from the comments collection.

	Categories of links		Contents of links	
Year	**Internal**	**External**	**Known media (%)**	**Images (%)**
2005	0	38	11 (28.95)	0 (0.0)
2006	1	19,051	4,679 (24.56)	556 (11.88)
2007	2	74,053	17,253 (23.3)	3,420 (19.82)
2008	12,299	325,930	74,095 (21.91)	23,315 (31.47)
2009	54,658	619,488	141,166 (20.94)	60,645 (42.96)
2010	99,215	1,169,786	280,332 (22.09)	154,372 (55.07)
2011	206,306	2,325,610	590,792 (23.33)	378,633 (64.09)
2012	370,074	4,513,511	1,317,284 (26.97)	970,935 (73.71)
2013	558,468	6,704,073	1,944,145 (26.77)	1,479,230 (76.09)
2014	779,987	9,283,846	2,575,111 (25.59)	1,988,173 (77.21)
2015	1,181,543	12,752,582	3,171,927 (22.76)	2,416,944 (76.2)
2016	1,630,462	16,470,491	3,538,101 (19.55)	2,577,966 (72.86)
2017	2,055,250	19,013,337	3,920,151 (18.61)	2,801,227 (71.46)
2018	1,974,185	14,021,587	2,702,087 (16.89)	1,840,742 (68.12)
2019	1,776,779	11,532,445	1,990,511 (14.96)	1,300,974 (65.36)
2020	533,875	3,333,823	536,813 (13.88)	323,796 (60.32)

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This publication was funded by Qatar Research Leadership program grant, ID# QRLP9-G-3330102, from the Qatar National Research Fund (a member of Qatar Foundation). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Qatar Research Leadership program grant from the Qatar National Research Fund (a member of Qatar Foundation): QRLP9-G-3330102.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Hind Almerekhi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Haewoon Kwak conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Bernard J. Jansen conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data is available at Github:
https://github.com/Hind-Almerekhi/toxicityChangesReddit.

REFERENCES

Alfonso F, Morris K. 2013. The most influential people on Reddit in 2013. *Available at https://www.dailydot.com/irl/reddit-top-10-2013-quickmeme-unidan-boston-bombers/* (accessed on 19 April 2019).

Almerekhi H, Kwak H, Jansen BJ. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In: *Companion proceedings of the web conference 2020*. New York, NY, USA: ACM, 294–298.

Ashraf N, Zubiaga A, Gelbukh A. 2021. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science* **7**:e742 DOI 10.7717/peerj-cs.742.

Badjatiya P, Gupta S, Gupta M, Varma V. 2017. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th international conference on world wide web companion*. Republic and Canton of Geneva, CHE. New York: ACM, 759–760.

Batista GEAPA, Prati RC, Monard MC. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **6(1)**:20–29 DOI 10.1145/1007730.1007735.

Baumgartner J. 2017. Directory contents. *Available at https://files.pushshift.io/reddit/* (accessed on 10 May 2021).

Bennett EM, Alpert R, Goldstein A. 1954. Communications through limited-response questioning. *Public Opinion Quarterly* **18(3)**:303–308 DOI 10.1086/266520.

Bowler L, Knobel C, Mattern E. 2015. From cyberbullying to well-being: a narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology* **66(6)**:1274–1293 DOI 10.1002/asi.23270.

Carton S, Mei Q, Resnick P. 2020. Feature-based explanations dont help people detect misclassifications of online toxicity. In: *Proceedings of the international AAAI conference on web and social media, volume 14*. Palo Alto: AAAI Press, 95–106.

Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E. 2017a. You can't stay here: the efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* **1**:31:1–31:22.

Chandrasekharan E, Samory M, Jhaver S, Charvat H, Bruckman A, Lampe C, Eisenstein J, Gilbert E. 2018. The internet's hidden rules: an empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* **2(CSCW)**:32 DOI 10.1145/3274301.

Chandrasekharan E, Samory M, Srinivasan A, Gilbert E. 2017b. The bag of communities: identifying abusive behavior online with preexisting internet data. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. New York, NY, USA: ACM, 3175–3187.

Cheng J, Bernstein M, Danescu-Niculescu-Mizil C, Leskovec J. 2017. Anyone can become a troll: causes of trolling behavior in online discussions. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. New York, NY, USA: ACM, 1217–1230.

Cheng J, Danescu-Niculescu-Mizil C, Leskovec J. 2015. Antisocial behavior in online discussion communities. In: *Proceedings of the 9th international conference on web and social media*. Palo Alto: AAAI Press, 61–70.

Choi D, Han J, Chung T, Ahn Y-Y, Chun B-G, Kwon TT. 2015. Characterizing conversation patterns in Reddit: from the perspectives of content properties and user participation behaviors. In: *Proceedings of the 2015 ACM on conference on online social networks*. New York, NY, USA: ACM, 233–243.

Davidson T, Warmsley D, Macy M, Weber I. 2017. Automated hate speech detection and the problem of offensive language. In: *Proceedings of the 11th international conference on web and social media*. Palo Alto: AAAI Press, 512–515.

Del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. 2017. Hate me, hate me not: hate speech detection on Facebook. In: *Proceedings of the 1st Italian conference on cybersecurity*. 86–95.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*. Minneapolis, Minnesota: ACL, 4171–4186.

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N. 2015. Hate speech detection with comment embeddings. In: *Proceedings of the 24th international conference on world wide web*. New York, NY, USA: ACM, 29–30.

Feinstein AR, Cicchetti DV. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* **43(6)**:543–549 DOI 10.1016/0895-4356(90)90158-L.

Fortuna P, Soler-Company J, Wanner L. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* **58(3)**:1–17 DOI 10.1016/j.ipm.2021.102524.

Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP. 2018. Convolutional neural networks for toxic comment classification. In: *Proceedings of the 10th hellenic conference on artificial intelligence*. New York, NY, USA: ACM, 35:1–35:6.

Gwet KL. 2014. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics, LLC.

Hanu L, Thewlis J, Haco S. 2021. How AI is learning to identify toxic online content. Scientific American. *Available at https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/* (accessed on 10 June 2022).

Hu X, Tang J, Zhang Y, Liu H. 2013. Social spammer detection in microblogging. In: *Proceedings of the 23rd international joint conference on artificial intelligence*. Palo Alto: AAAI Press, 2633–2639.

Jhaver S, Ghoshal S, Bruckman A, Gilbert E. 2018. Online harassment and content moderation: the case of blocklists. *ACM Transactions on Computer-Human Interaction* **25(2)**:1–33.

Johnson BG. 2018. Tolerating and managing extreme speech on social media. *Internet Research* **28(5)**:1275–1291 DOI 10.1108/IntR-03-2017-0100.

Johnson R, Zhang T. 2016. Supervised and semi-supervised text categorization using LSTM for region embeddings. In: *Proceedings of the 33rd international conference on international conference on machine learning*. 526–534.

Kaakinen M, Oksanen A, Rsnen P. 2018. Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior* **78**:90–97 DOI 10.1016/j.chb.2017.09.022.

Kapil P, Ekbal A. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems* **210**:106458 DOI 10.1016/j.knosys.2020.106458.

Kordyaka B, Jahn K, Niehaves B. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* **30(4)**:1081–1102 DOI 10.1108/INTR-08-2019-0343.

Kumar S, Hamilton WL, Leskovec J, Jurafsky D. 2018. Community interaction and conflict on the web. In: *Proceedings of the 2018 world wide web conference*. Republic and Canton of Geneva, CHE. New York: ACM, 933–943.

Kwon KH, Gruzd A. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* **27(4)**:991–1010 DOI 10.1108/IntR-02-2017-0072.

Lapidot-Lefler N, Barak A. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* **28(2)**:434–443 DOI 10.1016/j.chb.2011.10.014.

Le Q, Mikolov T. 2014. Distributed representations of sentences and documents. In: *Proceedings of the 31st international conference on international conference on machine learning*. 1188–1196.

Lin T-Y, Goyal P, Girshick R, He K, Dollár P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. Piscataway: IEEE, 2980–2988.

Madukwe KJ, Gao X. 2019. The thin line between hate and profanity. In: Liu J, Bailey J, eds. *AI 2019: advances in artificial intelligence*. Cham: Springer International Publishing, 344–356.

Maity SK, Chakraborty A, Goyal P, Mukherjee A. 2018. Opinion conflicts: an effective route to detect incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction* **2(CSCW)**:117 DOI 10.1145/3274386.

Massanari A. 2017. # Gamergate and The Fappening: how Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* **19(3)**:329–346 DOI 10.1177/1461444815608807.

Mathew B, Dutt R, Goyal P, Mukherjee A. 2019. Spread of hate speech in online social media. In: *Proceedings of the 10th ACM conference on web science*. New York, NY, USA: ACM, 173–182.

Mathew B, Illendula A, Saha P, Sarkar S, Goyal P, Mukherjee A. 2020. Hate begets hate: a temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* **4(CSCW2)**:92 DOI 10.1145/3415163.

Mittos A, Zannettou S, Blackburn J, De Cristofaro E. 2020. "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. In: *Proceedings of the international AAAI conference on web and social media, volume 14*. Palo Alto: AAAI Press, 452–463.

Mohan S, Guha A, Harris M, Popowich F, Schuster A, Priebe C. 2017. The impact of toxic language on the health of Reddit communities. In: *Proceedings of the 30th Canadian conference on artificial intelligence*. 51–56.

Mondal M, Silva LA, Benevenuto F. 2017. A measurement study of hate speech in social media. In: *Proceedings of the 28th ACM conference on hypertext and social media*. New York, NY, USA: ACM, 85–94.

Mukherjee P, Jansen BJ. 2017. Conversing and searching: the causal relationship between social media and web search. *Internet Research* **27(5)**:1209–1226 DOI 10.1108/IntR-07-2016-0228.

Newell E, Jurgens D, Saleem HM, Vala H, Sassine J, Armstrong C, Ruths D. 2016. User migration in online social networks: a case study on Reddit during a period of community unrest. In: *Proceedings of the 10th international conference on web and social media*. Palo Alto: AAAI Press, 279–288.

Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. 2016. Abusive language detection in online user content. In: *Proceedings of the 25th international conference on World Wide Web*. Republic and Canton of Geneva, CHE: ACM, 145–153.

Obadimu A, Khaund T, Mead E, Marcoux T, Agarwal N. 2021. Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. *Information Processing & Management* **58(5)**:1–18 DOI 10.1016/j.ipm.2021.102660.

Pelicon A, Shekhar R, krlj B, Purver M, Pollak S. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science* **7**:e559 DOI 10.7717/peerj-cs.559.

Perspective. 2017. Using machine learning to reduce toxicity online. *Available at https://www.perspectiveapi.com* (accessed on 30 November 2017).

Pronoza E, Panicheva P, Koltsova O, Rosso P. 2021. Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management* **58(6)**:1–23 DOI 10.1016/j.ipm.2021.102674.

Radfar B, Shivaram K, Culotta A. 2020. Characterizing variation in toxic language by social context. In: *Proceedings of the international AAAI conference on web and social media*. Palo Alto: AAAI.

Rafiq RI, Hosseinmardi H, Han R, Lv Q, Mishra S, Mattson SA. 2015. Careful what you share in six seconds: detecting cyberbullying instances in vine. In: *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining*. New York, NY, USA: ACM, 617–622.

Rajadesingan A, Resnick P, Budak C. 2020. Quick, community-specific learning: how distinctive toxicity norms are maintained in political subreddits. Palo Alto: AAAI Press, 557–568.

Rodriguez N, Rojas-Galeano S. 2018. Fighting adversarial attacks on online abusive language moderation. In: *Proceedings of the 5th workshop on engineering applications, volume 915*. Cham: Springer International Publishing, 480–493.

Salminen J, Almerekhi H, Kamel AM, Jung S-g, Jansen BJ. 2019. Online hate ratings vary by extremes: a statistical analysis. In: *Proceedings of the 2019 conference on human information interaction and retrieval*. New York: ACM, 213–217.

Salminen J, Almerekhi H, Milenković M, Jung S-g, An J, Kwak H, Jansen BJ. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: *Proceedings of the 12th international conference on web and social media*. Palo Alto: AAAI Press, 330–339.

Sazzed S. 2021. Identifying vulgarity in Bengali social media textual content. *PeerJ Computer Science* **7**:e665 DOI 10.7717/peerj-cs.665.

Shen Q, Rose C. 2019. The discourse of online content moderation: investigating polarized user responses to changes in Reddit's quarantine policy. In: *Proceedings of the third workshop on abusive language online*. Florence, Italy: ACL, 58–69.

Shores KB, He Y, Swanenburg KL, Kraut R, Riedl J. 2014. The identification of deviance and its impact on retention in a multiplayer Game. In: *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing*. New York, NY, USA: ACM, 1356–1365.

Silva LA, Mondal M, Correa D, Benevenuto F, Weber I. 2016. Analyzing the targets of hate in online social media. In: *Proceedings of the 10th international conference on web and social media*. AAAI Press, 687–690.

Singh S, Thapar V, Bagga S. 2020. Exploring the hidden patterns of cyberbullying on social media. *Procedia Computer Science* **167**:1636–1647 DOI 10.1016/j.procs.2020.03.374.

Squicciarini AC, Dupont J, Chen R. 2014. Online abusive users analytics through visualization. In: *Proceedings of the 23rd international conference on world wide web*. New York, NY, USA: ACM, 155–158.

Suler J. 2004. The online disinhibition effect. *Cyberpsychology & Behavior* **7(3)**:321–326 DOI 10.1089/1094931041291295.

Tsikerdekis M, Zeadally S. 2014. Online deception in social media. *Communications of the ACM* **57(9)**:72–80 DOI 10.1145/2629612.

Vaidya A, Mai F, Ning Y. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. *Proceedings of the International AAAI Conference on Web and Social Media* **14(1)**:683–693.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser u., Polosukhin I. 2017. Attention is all you need. In: *Proceedings of the 31st international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., 6000–6010.

Vens C, Costa F. 2011. Random forest based feature induction. In: *IEEE 11th international conference on data mining*. Piscataway: IEEE Computer Society, 744–753.

Vogels E. 2020. The state of online harassment. Pew Research. *Available at https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/* (accessed on 13 January 2021).

Wang X, Lee EW. 2021. Negative emotions shape the diffusion of cancer tweets: toward an integrated social network–text analytics approach. *Internet Research* **31(2)**:401–418.

Wulczyn E, Thain N, Dixon L. 2017. Ex machina: personal attacks seen at scale. In: *Proceedings of the 26th international conference on World Wide Web*. Republic and Canton of Geneva, CHE: ACM, 1391–1399.

Yin D, Xue Z, Hong L, Davison BD, Kontostathis A, Edwards L. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* **2**:1–7.

Yin W, Zubiaga A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* **7**:e598 DOI 10.7717/peerj-cs.598.

Zhang J, Danescu-Niculescu-Mizil C, Sauper C, Taylor SJ. 2018. Characterizing online public discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer Interaction* **2**:198 DOI 10.1145/3274467.

Zhao X, Liu JS, Deng K. 2013. Assumptions behind intercoder reliability indices. *Annals of the International Communication Association* **36(1)**:419–480 DOI 10.1080/23808985.2013.11679142.

Zhao Z, Zhang Z, Hopfgartner F. 2022. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media* **29**:1–14 DOI 10.1016/j.osnem.2022.100205.

Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In: *Proceedings of the 26th international conference on computational linguistics*. Osaka, Japan: ACL, 3485–3495.

PeerJ Computer Science

An optimized deep learning approach for suicide detection through Arabic tweets

Nadiah A. Baghdadi[1], Amer Malki[2], Hossam Magdy Balaha[3], Yousry AbdulAzeem[4], Mahmoud Badawy[3] and Mostafa Elhosseini[2,3]

[1] Nursing Management and Education Department, College of Nursing, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[2] College of Computer Science and Engineering, Taibah University, Yanbu, Saudi Arabia
[3] Computers and Control Systems Engineering Department, Faculty of Engineering, Mansoura University, Mansoura, Egypt
[4] Computer Engineering Department, Misr Higher Institute for Engineering and Technology, Mansoura, Egypt

ABSTRACT

Many people worldwide suffer from mental illnesses such as major depressive disorder (MDD), which affect their thoughts, behavior, and quality of life. Suicide is regarded as the second leading cause of death among teenagers when treatment is not received. Twitter is a platform for expressing their emotions and thoughts about many subjects. Many studies, including this one, suggest using social media data to track depression and other mental illnesses. Even though Arabic is widely spoken and has a complex syntax, depressive detection methods have not been applied to the language. The Arabic tweets dataset should be scraped and annotated first. Then, a complete framework for categorizing tweet inputs into two classes (such as Normal or Suicide) is suggested in this study. The article also proposes an Arabic tweet preprocessing algorithm that contrasts lemmatization, stemming, and various lexical analysis methods. Experiments are conducted using Twitter data scraped from the Internet. Five different annotators have annotated the data. Performance metrics are reported on the suggested dataset using the latest Bidirectional Encoder Representations from Transformers (BERT) and Universal Sentence Encoder (USE) models. The measured performance metrics are balanced accuracy, specificity, F1-score, IoU, ROC, Youden Index, NPV, and weighted sum metric (WSM). Regarding USE models, the best-weighted sum metric (WSM) is 80.2%, and with regards to Arabic BERT models, the best WSM is 95.26%.

Subjects Computational Linguistics, Network Science and Online Social Networks, Social Computing, World Wide Web and Web Science, Neural Networks
Keywords Deep Learning (DL), Machine Learning (ML), Suicide, Twitter

Submitted 9 May 2022
Accepted 28 July 2022
Published 23 August 2022

Corresponding author
Mahmoud Badawy,
engbadawy@mans.edu.eg

Academic editor
Muhammad Asif

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.1070**

Copyright
2022 Baghdadi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Millions of individuals worldwide suffer from various mental disorders. These disorders impact their thoughts and conduct, and their quality of living (*Kessler et al., 2017*; *Mathers & Loncar, 2006*). One of the most common mental illnesses globally is major depressive disorder (MDD), also known as depression. It has a significant influence on an individual's daily activities. It is linked to functional impairment, which is expensive to society. More than 300 million individuals are impacted globally, accounting for 4.4% of the global population, and the number is growing every day (*World Health Organization (WHO),*

2021). According to estimates, unipolar depression will contribute to the second-largest portion of the global illness burden by 2030. It is important to differentiate depression from the mood swings and transient emotions we experience every day (*Russell, 1980*). If it persists for longer than 2 weeks, it can become a serious health problem. It can impact both a person's mental and physical well-being. Thus, depressed people exhibit poor performance and behave unprofessionally at work and at home. If a depressed individual does not receive sufficient care, it might lead to suicide.

Every year, about 800,000 individuals commit suicide (*Safa, Bayat & Moghtader, 2022*). Among teenagers, it is the second leading cause of death. Depression affects millions of individuals worldwide, regardless of culture, gender, age, ethnicity, or economic condition. Symptoms of depression can be categorized psychologically, socially, and physically (*NHS, 2019*). The presence of all these symptoms is rare in depression patients, but they can predict the severity of the illness. Although a high percentage of depression is treated, over 70% of those affected do not receive therapy, which may cause their illness to progress (*Salas-Zárate et al., 2022*; *DBSA, 2022*). This emphasizes the importance of excellent mental health services and an innovative way to detect mental health disorders.

Traditional interactions between individuals and organizations have evolved. Meetings and conversations occur in online communities to exchange knowledge, obtain and provide assistance, and expand the business. Individuals use social media platforms such as Twitter and Facebook to express their emotions and thoughts about a wide range of topics (*Safa, Bayat & Moghtader, 2022*). An open and free communication channel such as social media facilitates problem resolution, and knowledge exchange (*Salas-Zárate et al., 2022*). The current status of society is deeply entwined with social networks. Last year, social media users climbed by around 424 million, bringing the overall number of users to 4.62 billion (*Kemp, 2022*). This equates to an annual increase of more than 10%. This evolution of social media enabled individuals to instantly express their feelings, ideas, emotions, and opinions. In the process, huge amounts of social data are generated, providing useful information about people's mental and everyday activities (*Wongkoblap, Vadillo & Curcin, 2022*). Mining and analyzing user data gives a unique chance to research human behavior in depth and their state of mental health and assist them in recovering.

Because depression influences human behavior, researchers devised a method for tracking depression and other mental diseases using social media data. The conventional study method for depression analysis relies on questionnaires, which involve participants providing subjective responses or comments. Social media data alone cannot provide people's true emotions since they differ from person to person and are hard to get. As a result, social media is frequently employed to detect stress, anxiety, and depression disorders. Research on psychology and human behavior can be conducted through social media sites like Twitter, and Facebook (*Mishra & Garg, 2018*; *Kim et al., 2020*). For example, studying tweets of individuals who have experienced severe depression has proven useful in predicting future depression in those individuals (*AlSagri & Ykhlef, 2020*; *Zafar & Chitnis, 2020*).

The analysis of social media data allows for the discreet detection of depressive symptoms before they proceed to more severe stages of depression. This enables the

recommendation of strategies for depression prevention and therapy in the early phases. Social media sites have been used to identify depression, with Twitter being the most commonly utilized (*Smys & Raj, 2021*; *Liu et al., 2022a*). Over 436 million monthly active Twitter users send more than 500 million Tweets each day, making Twitter the ninth most popular website on the Internet in terms of popularity (*Kemp, 2022*). Any registered user can publish his or her opinions in 140 characters at a time. Tweets are often made public and may be collected and analyzed using the Twitter API (*Liu et al., 2022b*). In addition, Twitter's API enables complicated searches such as retrieving every tweet regarding a specific topic (*Vanam & Jeberson, 2021*).

Speech and text analysis is carried out through natural language processing (NLP). During the development of computer technology, linguistic approaches evolved into computer algorithms. Natural language processing was originally used to examine classical tasks, such as structuring grammar in books. However, it has expanded to interpreting human viewpoints, including mail, online content, reviews, social networking postings, and media articles (*Zhang et al., 2022*). In many ways, NLP can process a variety of languages (*Islam et al., 2018*). For example, using sentiment analysis (SA), you can identify positive and negative aspects of a text or speech (*Bhushan & Sharma, 2021*). This method may be adapted to analyze a person's depression levels based on social media posts and negative sentiment ratings. Many prediction algorithms and optimization approaches exist for noticing patterns in data and forming insight based on those observations, such as deep learning (DL) and machine learning (ML). For example, when researchers use binary classification approaches for text analysis, they use traditional machine learning (ML) algorithms such as random forest (RF), support vector machines (SVMs), decision trees (DT), *etc*.

Understanding the available dataset can aid in determining whether machine learning or deep learning should be used to solve a particular problem. Machine learning is typically utilized when structured data are scarce. Most machine learning algorithms are built to train models using tabular data. Deep learning usually necessitates a significant amount of training data to ensure that the network, which may contain tens of millions of parameters, does not overfit the data. Hardware also plays an important role in choosing which AI method to use. For instance, machine learning algorithms use less computing power. Therefore, desktop computers are sufficient for training these types of algorithms. However, deep learning models require specialist hardware due to their high memory and computation requirements. To summarize, for deep learning to work, a tremendous amount of data must be collected, but minimal human intervention is required. Unfortunately, there is no cure for the shortage of large training datasets that transfer learning provides.However, the traditional ML technique's performance was restricted as the number of correlations rose dramatically due to the rise in data amount (*Rao et al., 2020*).

The use of deep learning (DL) in a variety of areas has been successfully demonstrated in recent years, including stock market forecasting (*Ni, Wang & Cheng, 2021*), traffic flow and accident risk prediction (*Essien et al., 2021*), and mental disease diagnosis (*Kim et al., 2020*). Deep learning has also been used to detect depression on social media, and the

results were significantly better than those achieved with traditional machine learning approaches. In addition, using Deep Learning models like RNNs and CNNs coupled with neural word embeddings has significantly improved textual data classification accuracy (*Sood et al., 2018*). This article presents a framework for analyzing texts (tweets) using six phases, starting with data acquisition and ending with classification, moving through four more steps. Finally, a trained and optimized model is used to classify the Arabic tweets into suicides or normal tweets. Accordingly, the following points outline the study's findings:

– Providing a comprehensive framework for classifying Arabic texts.
– Proposing a method of preprocessing tweets.
– Using the latest Arabic and universal models.
– Suggesting scraping and annotating the Arabic tweets dataset.
– Comparing different methods of lexical analysis, such as lemmatization and stemming.
– Reporting metrics of state-of-the-art performances from the Arabic tweet dataset.

What follows is a summary of the rest of the article: In "Related work", the related studies are presented. "Methodology" discusses the methodology in six phases and presents the suggested framework and algorithm. "Experiments and discussions" reports the experiments and their results and discussions. Finally, the article is concluded in "Conclusions", and the future work is presented therein.

RELATED WORK

Using social media data to detect depression has been the subject of many research works. Some of these research techniques are discussed here. *Victor et al. (2019)* developed multi-modal neural networks for detecting depression. Automatic evaluation is conducted using data collected from 671 participants. To evaluate depression, participants completed the Patient Health Questionnaire (PHQ-9) (*Kroenke, Spitzer & Williams, 2001*). This model uses artificial intelligence and the naive Bayes classification method to enhance system accuracy. They achieved 70.88% accuracy. In a recommendation system designed by *Rosa et al. (2018)*, CNN was used to identify users who may be suffering from psychological problems, such as depression and stress. One hundred forty-six assessors participated in a web-based questionnaire to build the system to identify their emotional states. The system delivers comforting messages to the user through 360 predefined messages. *Wang, Niu & Yu (2019)* developed SentiDiff, a novel algorithm based on the correlation between tweets and retweets. Combining this model with a traditional neural network resulted in a 79% accuracy. Nonetheless, the algorithm only considered tweets with retweets attached and only analyzed text-based information.

With CNN, *Akhtar et al. (2019)* achieved an accuracy of 80.52% in a single analysis, focusing on rough and fine-grained emotions and sentiment. They built a multi-task ensemble framework for emotion, sentiment, and intensity prediction. The framework employed a multi-layer perceptron network that learns from three deep learning models (*i.e.*, CNN, LSTM, and GRU) assisted by a handcrafted feature vector. The multi-task ensemble network achieved an accuracy of 89.88%. *Suman et al. (2020)* analyzed tweets for

sadness using DL models combined with a cloud-based app. This study employed RoBERTa to classify tweets using the linked tweets. The study achieved a testing accuracy of 87.23%. *Shetty et al. (2020)* employed SA to analyze Twitter tweets. For DL, KAGGLE datasets were used, and LSTM was used to build deep networks. CNNs were later proposed in the classification phase for improved performance. To discover the intensity of the sentiments of extremism, *Asif et al. (2020)* employed multinomial naive Bayes and linear support vector classifier algorithms on social media textual data. They classified the incorporated views into four categories based on their level of extremism. During the COVID-19 epidemic, *Alabdulkreem (2021)* adopted ML to analyze Arab women's tweets to determine whether they had depression symptoms. From 10,000 tweets taken from 200 individuals, they used an RNN approach to predict depression. They achieved an accuracy of 72%.

According to *Chiu et al. (2021)*, using Instagram photos, text, and behavior characteristics of posts, a multi-model architecture is designed to predict depression based on the DL. They applied an aggregation method with a two-stage detection mechanism to recognize the hidden emotions of people. *Tommasel et al. (2021)* analyzed social media expression in Argentina as part of their DL approach. Time-series data were generated (using markers) and entered into a neural network during the COVID-19 pandemic to predict psychological health and feelings. In a study by *Jyothi Prasanth, Dhalia Sweetlin & Sruthi (2022)*, they developed a method to diagnose depression among Twitter users by reducing multiclassification rates. An improved method for classifying negative tweets is proposed. A glossary of emoticons is created, and genuinely negative tweets are detected. They achieved accuracies of 72%, 76%, and 90% for RNN, LSTM, and BiLSTM networks, respectively. *Zogan et al. (2022)* developed explainable Multi-aspect depression detection utilizing social media analysis to detect depressed people. It employs a hierarchical attention network to extract users' online activity and behavior characteristics. They also compared the proposed technique with both traditional and deep learning techniques. Their system achieved 89.5% accuracy. *Kour & Gupta (2022)* developed a unique way of predicting depression using raw Twitter datasets by combining CNN and biLSTM networks. To analyze longer text sequences, the proposed model employs biLSTM. *Nair et al. (2022)* employed machine learning techniques to predict early signs of depression. Classifiers such as naive Bayes, decision trees, SVM, and k-nearest neighbors are employed on big data extracted from social media. The decision trees classifier achieved the highest results with 97% accuracy. Using clinical interviews, *Park & Moon (2022)* built a multi-modal system to detect depression. They used BERT-CNN for natural language analysis and CNN-BiLSTM for voice signal processing. They proved that a fusion of text data and voice data improves detection accuracy.

Safa, Bayat & Moghtader (2022) proposed a multi-modal framework to predict depression symptoms from Twitter profiles. They applied lexicon analysis and NLP techniques to study the relationship between language and depression. In addition, they explored new features, such as bio-text, profile picture, and header image. To identify image content, a convolutional neural network-based automatic tagging system is used to tag both profile and header images. Their results showed that the correlation-based

method leads to better outcomes. *Lia et al. (2022)* used six generic machine learning algorithms to detect depression from Twitter's tweets. They used term frequency-inverse document frequency for feature extraction before applying classification algorithms. Among the used techniques, the support vector classifier outperforms all the other techniques with 79.90% accuracy. *Shankdhar, Mishra & Shukla (2022)* used natural language processing tools and deep learning algorithms to predict depression through users' tweets. First, they built a baseline model using LSTM and achieved an accuracy of 95.12%. Afterward, they enhanced the model using a hybrid BiLSTM+CNN model. BiLSTM layers extract superior sequence features from text data, whereas CNN improves classification with higher dimensional features. To identify depression early on, *Tong et al. (2022)* proposed a new classifier, namely, cost-sensitive boosting pruning trees. They used their classifier on five distinct datasets, including two publicly available Twitter depression detection datasets and three datasets from the UCI machine learning repository. The proposed classifier is compared to state-of-the-art Boosting algorithms to assess it. In addition, they merged the Discrete Adaboost structure with the pruning procedure (Adaboost+PT) as a comparison approach. The proposed classifier outperformed the previous depression detection methods in the two Twitter datasets.

Table 1 provides a brief comparison of the techniques under consideration. Language, data source, data size, and achieved accuracy are stated for each technique. As demonstrated by the comparison, there is a lack of depression detection approaches applied to the Arabic language, despite its widespread use and difficult syntax. There is a crucial need for an empirical quantitative framework for text classification and depression detection for Arabic.

METHODOLOGY

The current study suggests a framework to perform text classification in six phases. It is summarized graphically in Fig. 1 and discussed in the following subsections. In practice, the Arabic tweet is classified using the optimized and trained model, and the output is generated to be "Suicide" or "Normal" as shown in Fig. 2. It tried to evaluate seven tweets that predicted six correctly and one incorrectly.

Phase 1: data acquisition

The authors scraped the data from Twitter using the "Twint" Python package (*TWINT, 2022*). Twint is an advanced Twitter scraping and OSINT tool. The authors worked on scraping six keywords (with and without hashtags) that have the native meaning of the English word "Suicide". They are listed in Table 2. The authors tried to retrieve the tweets from "2017-01-01" to "2022-01-20" (*i.e.*, when the script was running). The authors retrieved 14,576 tweets concerning the search keywords and specified duration. Table 3 show five samples from the retrieved records.

Phase 2: data pre-processing

With an initial investigation, after the scraping phase, multiple tweets are duplicated or meaningless (*i.e.*, containing only keywords, links, and/or emojis). This can be considered

Table 1 A comparison between state-of-the-art depression detection techniques.

Paper	Year	Technique	Language	Data source	Data size	Accuracy (%)
Victor et al. (2019)	2019	BiLSTM	English	Questionnaire	671 participants	70.88
Rosa et al. (2018)	2019	CNN+BiLSTM	Portuguese	Questionnaire	146 participants	90
Wang, Niu & Yu (2019)	2020	Sentiment diffusion patterns	English	Twitter	100,000 tweets	79
Akhtar et al. (2019)	2020	CNN+LSTM+GRU	English	EmoInt	7,102	89.88
				EmoBank	10,062	
				Facebook post	2,895	
				SemEval-2016	28,632	
Suman et al. (2020)	2020	DL	English	Twitter	1,600,000	87.23
Shetty et al. (2020)	2020	LSTM	English	Twitter	N/A	70
		Model vector				76.69
Alabdulkreem (2021)	2021	RNN-LSTM	Arabic	Twitter	10,000	72
Chiu et al. (2021)	2021	CNN for image	English	Instagram	520 users	N/A
		Word2vec for text				
		Handcrafted features for behavior				
Tommasel et al. (2021)	2021	RNN	Spanish	Twitter	150 million tweets	N/A
Jyothi Prasanth, Dhalia Sweetlin & Sruthi (2022)	2022	RNN	English	Twitter	1,200 users	72
		LSTM				76
		BiLSTM				90
Zogan et al. (2022)	2022	CNN+RNN	English	Twitter	4,208 users	89.5
Kour & Gupta (2022)	2022	CNN-biLSTM	English	Twitter	1,402 depressed	94.28
					300 million non-depressed	
Nair et al. (2022)	2022	Machine learning	English	Twitter	10,000 tweets	97
Park & Moon (2022)	2022	BERT-CNN	English	DAIC-WOZ	N/A	N/A
Safa, Bayat & Moghtader (2022)	2022	Machine learning	English	Twitter	11,890,632 tweets	91
Lia et al. (2022)	2022	Machine learning	English	Twitter	1,600,000	79.9
Shankdhar, Mishra & Shukla (2022)	2022	BiLSTM+CNN	English	Twitter	12,274 tweets	95.12
Tong et al. (2022)	2022	Discrete Adaboost Cost-sensitive Boosting Pruning Trees	English	TTDD	7,873 tweets	88.39
				CLPsych 2015	1,746 tweets	70.69
				LSVT	128	85.72
				Statlog	6,435	92.21
				Glass	214	77.63

noise in the data; hence, a pre-processing phase is necessary for the current scenario. Different pre-processing steps are applied as shown in Fig. 3. The steps include: (1) the scraped tweets are read, (2) the duplicated tweets are eliminated, (3) the suggested Arabic

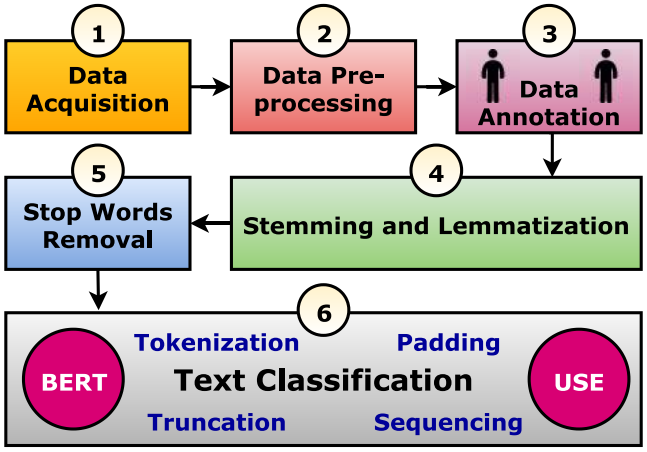


Figure 1 The suggested text classification framework. Full-size DOI: 10.7717/peerj-cs.1070/fig-1

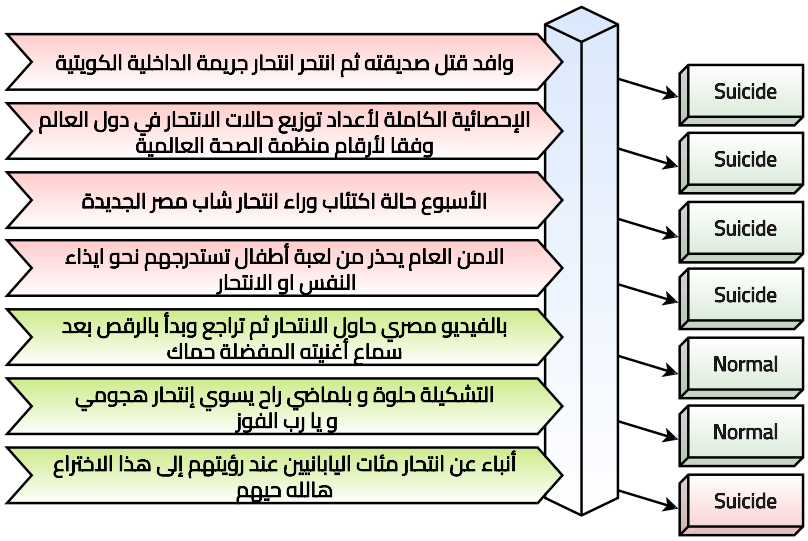


Figure 2 System evaluation in practice. Full-size DOI: 10.7717/peerj-cs.1070/fig-2

Table 2 The six scraped keywords (with and without hashtags).

Sr.	Without hashtag	With hashtag
1	انتحار	# انتحار
2	الإنتحار	# الإنتحار
3	الانتحار	# الانتحار
4	إنتحار	# إنتحار
5	منتحر	# منتحر
6	سانتحر	# سانتحر

RegEx pre-processing algorithm (discussed in the following paragraph) is applied, (4) the empty and duplicated tweets are eliminated, and (5) the tweets are sorted in ascending order concerning the create timestamp (*i.e.*, data, time, and timezone).

Table 3 Five samples from the scraped tweets.

Sr.	Tweet
1	إنتحار سياسي
2	التشكيلة حلوة و بلماضي راح يسوي إنتحار هجومي و يا رب الفوز
3	محاولة انتحار سيا بعد انتقادات لفيلمها
4	انتحار شاب في الناصرية حي اور
5	انتحار شابة بسبب دواء ضد حب الشباب

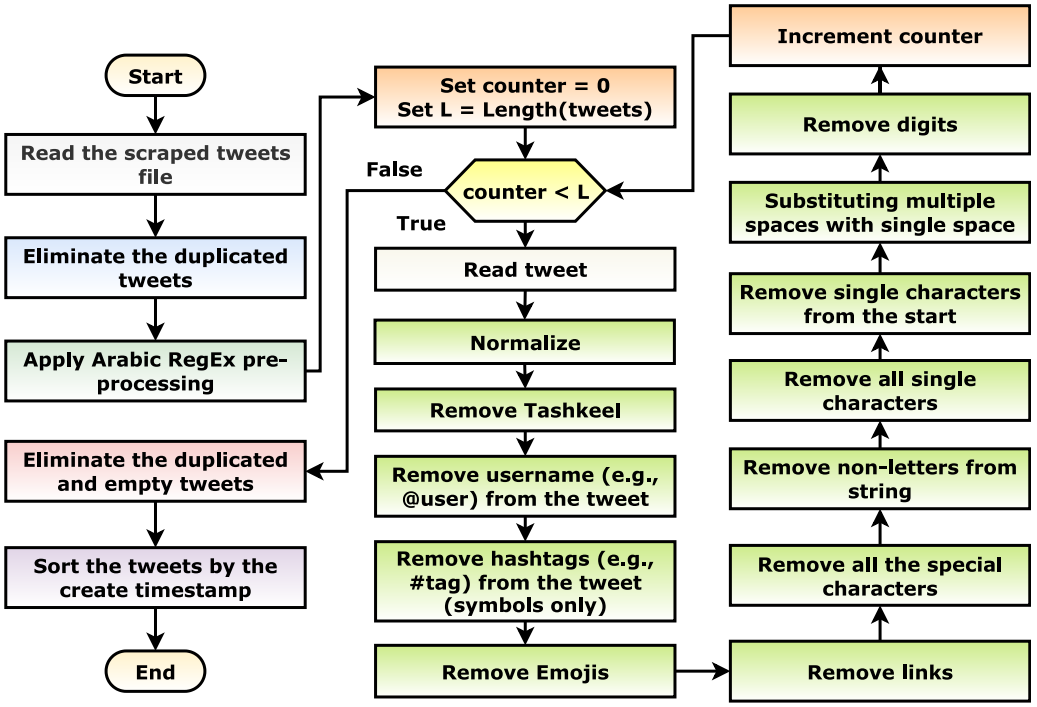


Figure 3 The tweets pre-processing flowchart. Full-size DOI: 10.7717/peerj-cs.1070/fig-3

The suggested Arabic regular expression pre-processing algorithm iterates on each tweet and apply: (1) normalize the tweet by changing a set of characters to its origin character (Table 4), (2) remove the Tashkeel (*i.e.*, "Shadda", "Fatha", "Tanwin Fath", "Damma", "Tanwin Damm", "Kasra", "Tanwin Kasr", "Sukun", and "Tatwil/Kashida"), (3) remove the username (*e.g.*, @hossam) using the "`@[^\s]+`" regular expression, (4) remove the hashtag symbol (*i.e.*, "#"), (5) remove the emojis, (6) remove the links using the "`(?:\@|http?\://|https?\://|www)\S+`" and "`((www\.[^\s]+)|(https?://[^\s]+))`" regular expressions, (7) remove all of the special characters using the "\W" regular expression, (8) remove the non-letters using the "`[^0-9\u0600-\u06FF]+`" regular expression, (9) remove all of the single characters using the "`\s+[0-9\u0600-\u06FF]\s+`" regular expression, (10) remove the single characters from the beginning of the tweet using the "`\^[0-9\u0600-\u06FF]\s+`" regular expression, (11) substitute the

Table 4 The normalized characters.

Sr.	From character(s)	To character
1	إ أ آ ا	ا
2	ي	ى
3	ة	ه
4	گ	ك
5	ؤ ئ	ء

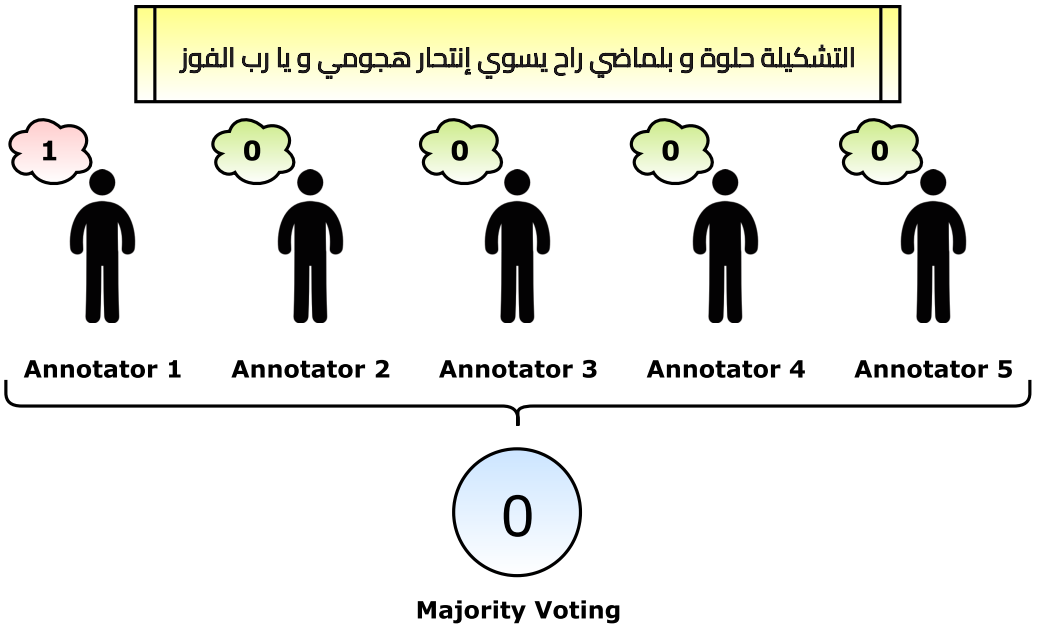


Figure 4 Graphical summarization of the voting process.
Full-size DOI: 10.7717/peerj-cs.1070/fig-4

multiple white spaces with single ones using the "`\s+`" and "`+`" regular expressions, and finally, (12) remove the digits using the "`[0-9]+`" regular expression.

Phase 3: data annotation

After applying the pre-processing steps, the number of remaining and unique tweets is 2,030. The next phase is to annotate the dataset. Five annotators helped in the annotation phase. Each annotator retrieved an empty sheet with the tweets only. He/she read the tweets one by one and marked "1" for the suicide tweets and "0" otherwise. The criteria for marking a suicide tweet is the tweet context. In other words, if the real contextual meaning of the tweet shows suicide, then it is marked by "1". The reason is that some tweets reflect some different meanings, such as sarcasm (*e.g.*, the fifth row in Table 3). After that, the majority voting algorithm is applied to gain the correct annotations. For example, if a tweet had [1, 0, 0, 1, 1], then it is considered a suicide tweet because the number of ones is more than zeros. Figure 4 reflects a graphical summarization of the voting process. After the annotation process, 1,074 tweets are categorized as "Normal" and 956 are categorized as

Table 5 Six samples from the processed and annotated tweets.

Sr.	Tweet	Category
1	وافد قتل صديقته ثم انتحر انتحار جريمة الداخلية الكويتية	1
2	الإحصائية الكاملة لأعداد توزيع حالات الانتحار في دول العالم وفقا لأرقام منظمة الصحة العالمية	1
3	الأسبوع حالة اكتئاب وراء انتحار شاب مصر الجديدة	1
4	الامن العام يحذر من لعبة أطفال تستدرجهم نحو ايذاء النفس او الانتحار	1
5	بالفيديو مصري حاول الانتحار ثم تراجع وبدأ بالرقص بعد سماع أغنيته المفضلة حماك	0
6	أنباء عن انتحار مئات اليابانيين عند رؤيتهم إلى هذا الاختراع هالله حيهم	0

Table 6 A sample comparison between a stemmed and lemmatized tweet.

Type	Tweet
Before Normalization	شاب ينتحر بسبب ما فعلته مع أسرته ورسالته قبل الانتحار بدقائق أنا رايح لربنا يجيبلي حقي صور
After Normalization	شاب ينتحر بسبب ما فعلته مع اسرته ورسالته قبل الانتحار بدقاءق انا رايح لربنا يجيبلي حقي صور
Stemmed	شاب نحر سبب ما فعل مع اسر رسل قبل نحر دقءق انا ريح لرب جيبل حقي صور
Lemmatized	شاب انتحر سبب ما فعل مع اسرته رسالة قبل انتحار بدقاءق انا رايح رب يجيبلي حق صور

"Suicide". Samples from the annotated and processed tweets are shown in Table 5. In the future, to overcome this issue for Arabic tweets, automatic learning can be applied as follows: (1) the built system can accept the future tweets and predict them, (2) volunteer annotators can determine whether the system predicted correctly or wrongly, and (3) the new tweet can be used to update the model weights.

Phase 4: stemming and lemmatization

The stemming and lemmatization processes are applied to the pre-processed and annotated tweets. The target of stemming is to remove (*i.e.*, stem) the last few characters of a certain word. It can lead to incorrect meanings and spelling. On the other hand, lemmatization considers the context and converts the word to its meaningful base form (*i.e.*, lemma). The Information Science Research Institute's (ISRI) Arabic Stemmer (*Taghva, Elkhoury & Coombs, 2005*) and Qalsadi Arabic Lemmatizer (*Zerrouki, 2012*) are used to perform that. A stemmed and lemmatized tweet is shown in Table 6.

Phase 5: stop words removal

After the stemming and lemmatization phase, the stop words are removed from the tweets. The current study worked on 754 stop words in Arabic (*NLTK, 2022*). The process is to scan each token (*i.e.*, single word) in the tweet, if it belongs to the stop words, then it is eliminated. Examples of the used stop words are: "شرع" "كاد", and "مازال". In the current phase, there are six different datasets flavours to compare with. They are (1) the tweets before normalization, (2) the tweets after normalization, (3) the stemmed and normalized tweets, (4) the lemmatized and normalized tweets, (5) the filtered stemmed and normalized tweets, and (6) the filtered lemmatized and normalized tweets.

Phase 6: classification

The data consists of tweets and the corresponding labels. A tweet vector can be though a distributed representation of that tweet and computed for every tweet in the dataset. The tweet vector and the corresponding label are used to train and validate the statistical classification model. The intuition is that the tweet vector captures the tweet semantics and hence can be effectively used for classification. Two alternatives, common for several text classification problems in NLP, can be used to obtain the tweets vectors. They are (1) word-based and (2) context-based representations. The word-based representation combines word embeddings of the tweet content words, and then the average of the word embeddings is calculated. The context-based representation can handle language models to generate vectors of the sentences. So, instead of learning vectors for individual words in the sentence as happens with the word-based representation, that latter representation computes a vector for sentences as a whole by considering the order of words and the set of co-occurring words.

GloVe and Word2Vec are examples of the first representation, while Embeddings from Language Models (ELMo), Neural-Net Language Model (NNLM), Bidirectional Encoder Representations from Transformers (BERT), and Universal Sentence Encoder (USE) are examples on the second representation. *Joshi et al. (2019)* recommend the latter approach than the first approach. The current study utilizes the USE and BERT with the Arabic flavors to perform the text classification task. The BERT uses the transformer architecture besides a set of different techniques to train a model, resulting in a state-of-the-art performance. The USE also uses the transformer architecture, not the recurrent neural networks (RNNs), similar to the ELMo approach.

For the USE, the universal-sentence-encoder-multilingual (*Google, 2019a*) and universal-sentence-encoder-multilingual-qa (*Google, 2019b*) are used. They support the Arabic language alongside another 15 languages. The used architecture is built as follows (1) the USE embedding layer, (2) a dense layer with 128 neurons and ReLU activation function, (3) another dense layer with 64 neurons and ReLU activation function, and (4) an output activation function with the SoftMax activation function. The embedding size is set to 256. The optimizer is set to Adam with a learning rate of 0.0001. The loss function is set to "Categorical Crossentropy".

For the BERT, the Arabic BERT versions: AraBERT (*Antoun, Baly & Hajj, 2020*) and AraELECTRA (*Antoun, Baly & Hajj, 2021*) are used. Five different flavors from them are used: "bert-base-arabertv2", "araelectra-base-generator", "bert-base-arabertv02-twitter", "bert-large-arabertv02-twitter", and "bert-large-arabertv2". Intuitively, using BERT models requires additional pre-processing as the model does not perform this automatically. Hence, the sentences are tokenized, truncated, padded, and converted to a sequence before being classified. The max sequence length is set to 100.

Table 7 summarizes the used models and their configurations. The last column defines a keyword to be used in the experiments instead of using the whole model name. To evaluate the model, different performance metrics are calculated (*Seliya, Khoshgoftaar & Van Hulse, 2009*; *Baghdadi et al., 2022b*, *2022a*): (1) accuracy (Eq. (1)), (2) balanced accuracy

Table 7 Summary on the used models and the their configurations.

Approach	Model name	Size	Pre-process?	Experiment keyword
USE	universal-sentence-encoder-multilingual	245.48 MB	No	USE-M
USE	universal-sentence-encoder-multilingual-qa	310.68 MB	No	USE-MQA
USE	universal-sentence-encoder-multilingual (Tuned)	245.48 MB	No	USE-M (Tuned)
USE	universal-sentence-encoder-multilingual-qa (Tuned)	310.68 MB	No	USE-MQA (Tuned)
BERT	bert-base-arabertv2	543 MB	Yes	BERT-BV2
BERT	bert-large-arabertv2	1.38 GB	Yes	BERT-LV2
BERT	araelectra-base-generator	227 MB	Yes	AraElectra
BERT	bert-base-arabertv02-twitter	543 MB	Yes	BERT-BV02T
BERT	bert-large-arabertv02-twitter	1.38 GB	Yes	BERT-LV02T

(Eq. (2)), (3) precision (Eq. (3)), (4) specificity (Eq. (4)), (5) recall (*i.e.*, sensitivity) (Eq. (5)), (6) F1-score (Eq. (6)), (7) IoU (Eq. (7)), (8) ROC (Eq. (8)), (9) Youden Index (Eq. (9)), and (10) NPV (Eq. (10)). In addition to them, a weighted sum metric (WSM) is calculated mentioned metrics using Eq. (11).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Balanced Accuracy} = 0.5 \times (\text{Recall} + \text{Specificity}) \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{ROC} = \frac{\sqrt{\text{Sensitivity}^2 + \text{Specificity}^2}}{\sqrt{2}} \tag{8}$$

$$\text{Youden Index} = \text{Specificity} + \text{Sensitivity} - 100\% \tag{9}$$

$$\text{NPV} = \frac{TN}{TN + FN} \tag{10}$$

$$\begin{aligned}\text{WSM} = \frac{1}{10} \times (&\text{Accuracy} + \text{Balanced Accuracy} + \text{Precision} + \\ &\text{Specificity} + \text{Recall} + \text{F1} + \text{IoU} + \text{ROC} + \text{Youden Index} + \text{NPV})\end{aligned} \tag{11}$$

The suggested algorithm pseudocode

Algorithm 1 summarizes the different phases for fitting and testing a model. It is summarized graphically in Fig. 5.

Algorithm 1 **The algorithm pseudocode for fitting and testing a model.**

```
// Load the target dataset (one of the different created datasets in Subsection 3).
tweets, labels = LoadDataset()
// Create a tokenizer and fit the tweets on it.
tokenizer = CreateFitTokenizer()
tokens = tokenizer. fit(tweets)
// Convert the tweets to sequences.
sequences = ConvertToSequences(tokens)
// Pad the sequences with a predefined max length, post padding, and post truncating.
paddedSequences = PadAndTruncate(sequences)
// Convert the labels to categorical labels.
catLabels = LabelsToCategoricalLabels(labels)
// Create the model and initialize it. The model can be one of the model discussed in
  Table 7
model = CreateModel()
// Split the dataset into training, testing, and validation with a predefined train-to-
  test split percentage.
trainX, trainY, testX, testY, valX, valY = SplitDataset(paddedSequences, catLabels)
// Train and fit the created model on the datasets for a predefined epochs.
model.fit(trainX, trainY, valX, valY)
// Test the trained model using the testing subset and report the different performance
  metrics.
metrics = model.evaluate(testX, testY)
```

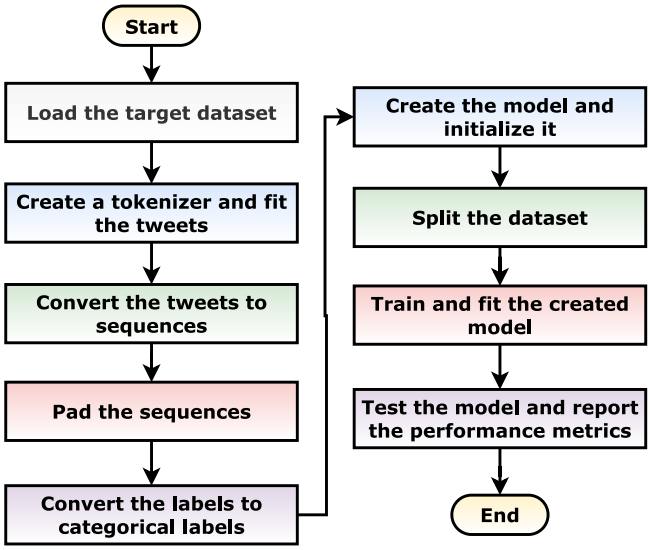


Figure 5 **Graphical flowchart summarization of the suggested algorithm.**
Full-size DOI: 10.7717/peerj-cs.1070/fig-5

Table 8 The experiments configurations.

Configuration	Specifications
Datasets	Subsection 3
Sequence Max Length	100
No. of Epochs	25
Batch size	32
Train-to-test Size	85% to 15%
Models	Table 7
No. of USE Models	3
No. of BERT Models	5
Scripting Language	Python
Working Environment	Google Colab (*i.e.*, Intel(R) Xeon(R) CPU @ 2.00 GHz, Tesla T4 16 GB GPU, CUDA v.11.2, and 12 GB RAM)

EXPERIMENTS AND DISCUSSIONS

This section presents the experiments and the discussions of the reported results. The section is partitioned into two sections (1) experiments applied using the USE models and (2) experiments applied using the BERT models. Table 8 summarizes the common configurations of all experiments.

The USE experiments

The current subsection discussed the applied experiments concerning the four USE flavors. Table 9 shows the loss and confusion matrix results of the four USE flavors (Table 7) on the six datasets (as listed in Subsection 3). From Table 9, the different discussed performance metrics can be calculated. They are reported in Table 10. From Table 9, it shows that the lowest loss of the different USE models is (1) 0.431 which is reported by "USE-MQA" using the "Before Normalization" dataset, (2) 0.473 which is reported by "USE-MQA" using the "After Normalization" dataset, (3) 0.526 which is reported by "USE-MQA" using the "Stemmed" dataset, (4) 0.477 which is reported by "USE-MQA" using the "Lemmatized" dataset, (5) 0.570 which is reported by "USE-M" using the "Stemmed and Filtered" dataset, and (6) 0.539 which is reported by "USE-MQA" using the "Lemmatized and Filtered" dataset. From Table 10, it shows that the highest WSM value of the different USE models is (1) 80.20% using the "Before Normalization" dataset, (2) 74.35% using the "After Normalization" dataset, (3) 77.47% using the "Stemmed" dataset, (4) 73.13% using the "Lemmatized" dataset, (5) 71.39% using the "Stemmed and Filtered" dataset, and (6) 70.59% using the "Lemmatized and Filtered" dataset. The best overall WSM value is 80.20%, produced by the "USE-MQA" model and "Before Normalization" dataset. The lowest overall WSM value is 61.06%, produced by the "USE-MQA" model and "Stemmed and Filtered" dataset. The average WSM value is 71.42%. Table 11 summarizes the WSM metrics and presents the maximum values of each row and column. Also, they are summarized graphically in Fig. 6.

Table 9 The loss and confusion matrix results of the four USE flavors on the six datasets.

Keyword	Dataset	Loss	TP	TN	FP	FN
USE-M (Tuned)	Before Normalization	0.718	109	136	25	35
USE-MQA (Tuned)	Before Normalization	0.442	116	136	25	28
USE-M	Before Normalization	0.444	123	127	34	21
USE-MQA	Before Normalization	0.431	111	144	17	33
USE-M (Tuned)	After Normalization	1.063	111	118	43	33
USE-MQA (Tuned)	After Normalization	1.338	114	118	43	30
USE-M	After Normalization	0.538	117	112	49	27
USE-MQA	After Normalization	0.473	115	123	38	29
USE-M (Tuned)	Stemmed	0.750	115	132	29	29
USE-MQA (Tuned)	Stemmed	1.590	103	126	35	41
USE-M	Stemmed	0.527	113	111	50	31
USE-MQA	Stemmed	0.526	94	130	31	50
USE-M (Tuned)	Lemmatized	0.483	90	142	19	54
USE-MQA (Tuned)	Lemmatized	1.344	102	123	38	42
USE-M	Lemmatized	0.523	113	114	47	31
USE-MQA	Lemmatized	0.477	112	123	38	32
USE-M (Tuned)	Stemmed and Filtered	1.039	106	125	36	38
USE-MQA (Tuned)	Stemmed and Filtered	1.189	93	120	41	51
USE-M	Stemmed and Filtered	0.570	100	115	46	44
USE-MQA	Stemmed and Filtered	0.598	68	139	22	76
USE-M (Tuned)	Lemmatized and Filtered	1.102	101	123	38	43
USE-MQA (Tuned)	Lemmatized and Filtered	0.877	103	123	38	41
USE-M	Lemmatized and Filtered	0.618	104	121	40	40
USE-MQA	Lemmatized and Filtered	0.539	114	113	48	30

The USE Experiments Remarks: It can be conducted from Tables 9–11, and Fig. 6 that (1) the lemmatization is better than stemming concerning all non-tuned USE models without filtering and the "USE-M (Tuned)" model with filtering and the stemming is better than lemmatization otherwise, (2) the lemmatization without filtering is better than the lemmatization with filtering concerning all USE models unless the "USE-MQA (Tuned)" model, (3) the stemming without filtering is better than the stemming with filtering concerning all USE models, and (4) there are major differences between using the normalization and ignoring it where the ignorance is better.

The BERT experiments

The current subsection discussed the applied experiments concerning the five BERT flavors. Table 12 shows the loss and confusion matrix results of the five BERT flavors (Table 7) on the six datasets (as listed in Subsection 3). The different discussed performance metrics can be calculated from Table 12. They are reported in Table 13. From Table 12, it shows that the lowest loss of the different BERT models is (1) 0.443 which is

Table 10 The performance metrics of the four USE flavors on the six datasets.

Keyword	Dataset	Accuracy (%)	Balanced accuracy (%)	Precision (%)	Specificity (%)	Recall (%)	F1 (%)	IoU (%)	ROC (%)	Youden index (%)	NPV (%)	WSM (%)
USE-M (Tuned)	Before Normalization	80.33	80.08	81.34	84.47	75.69	78.42	64.50	80.20	60.17	79.53	76.47
USE-MQA (Tuned)	Before Normalization	82.62	82.51	82.27	84.47	80.56	81.40	68.64	82.54	65.03	82.93	79.30
USE-M	Before Normalization	81.97	82.15	78.34	78.88	85.42	81.73	69.10	82.21	64.30	85.81	78.99
USE-MQA	Before Normalization	83.61	83.26	86.72	89.44	77.08	81.62	68.94	83.49	66.52	81.36	80.20
USE-M (Tuned)	After Normalization	75.08	75.19	72.08	73.29	77.08	74.50	59.36	75.21	50.38	78.15	71.03
USE-MQA (Tuned)	After Normalization	76.07	76.23	72.61	73.29	79.17	75.75	60.96	76.29	52.46	79.73	72.25
USE-M	After Normalization	75.08	75.41	70.48	69.57	81.25	75.48	60.62	75.63	50.82	80.58	71.49
USE-MQA	After Normalization	78.03	78.13	75.16	76.40	79.86	77.44	63.19	78.15	56.26	80.92	74.35
USE-M (Tuned)	Stemmed	80.98	80.92	79.86	81.99	79.86	79.86	66.47	80.93	61.85	81.99	77.47
USE-MQA (Tuned)	Stemmed	75.08	74.89	74.64	78.26	71.53	73.05	57.54	74.97	49.79	75.45	70.52
USE-M	Stemmed	73.44	73.71	69.33	68.94	78.47	73.62	58.25	73.86	47.42	78.17	69.52
USE-MQA	Stemmed	73.44	73.01	75.20	80.75	65.28	69.89	53.71	73.42	46.02	72.22	68.29
USE-M (Tuned)	Lemmatized	76.07	75.35	82.57	88.20	62.50	71.15	55.21	76.44	50.70	72.45	71.06
USE-MQA (Tuned)	Lemmatized	73.77	73.62	72.86	76.40	70.83	71.83	56.04	73.67	47.23	74.55	69.08
USE-M	Lemmatized	74.43	74.64	70.63	70.81	78.47	74.34	59.16	74.74	49.28	78.62	70.51
USE-MQA	Lemmatized	77.05	77.09	74.67	76.40	77.78	76.19	61.54	77.09	54.18	79.35	73.13
USE-M (Tuned)	Stemmed and Filtered	75.74	75.63	74.65	77.64	73.61	74.13	58.89	75.65	51.25	76.69	71.39
USE-MQA (Tuned)	Stemmed and Filtered	69.84	69.56	69.40	74.53	64.58	66.91	50.27	69.74	39.12	70.18	64.41
USE-M	Stemmed and Filtered	70.49	70.44	68.49	71.43	69.44	68.97	52.63	70.44	40.87	72.33	65.55
USE-MQA	Stemmed and Filtered	67.87	66.78	75.56	86.34	47.22	58.12	40.96	69.58	33.56	64.65	61.06
USE-M (Tuned)	Lemmatized and Filtered	73.44	73.27	72.66	76.40	70.14	71.38	55.49	73.33	46.54	74.10	68.67
USE-MQA (Tuned)	Lemmatized and Filtered	74.10	73.96	73.05	76.40	71.53	72.28	56.59	74.00	47.93	75.00	69.48
USE-M	Lemmatized and Filtered	73.77	73.69	72.22	75.16	72.22	72.22	56.52	73.70	47.38	75.16	69.20
USE-MQA	Lemmatized and Filtered	74.43	74.68	70.37	70.19	79.17	74.51	59.38	74.81	49.35	79.02	70.59

reported by "BERT-BV2" using the "Lemmatized" dataset, (2) 0.371, which is reported by "BERT-LV2" using the "Lemmatized and Filtered" dataset, (3) 0.384 which is reported by "AraElectra" using the "Before Normalization" dataset, (4) 0.330 which is reported by "BERT-BV02T" using the "Before Normalization" dataset, and (5) 0.404 which is reported by "BERT-LV02T" using the "After Normalization" dataset. From Table 13, it shows that

Table 11 The WSM metrics of the four USE flavors on the six datasets.

	Before normalization (%)	After normalization (%)	Stemmed (%)	Lemmatized (%)	Stemmed and filtered (%)	Lemmatized and filtered (%)	MAX (%)
USE-M (Tuned)	76.47	71.03	77.47	71.06	71.39	68.67	77.47
USE-MQA (Tuned)	79.30	72.25	70.52	69.08	64.41	69.48	79.30
USE-M	78.99	71.49	69.52	70.51	65.55	69.20	78.99
USE-MQA	**80.20**	74.35	68.29	73.13	61.06	70.59	**80.20**
MAX	**80.20**	74.35	77.47	73.13	71.39	70.59	**80.20**

Note:
The bold results reflect the best reported results.

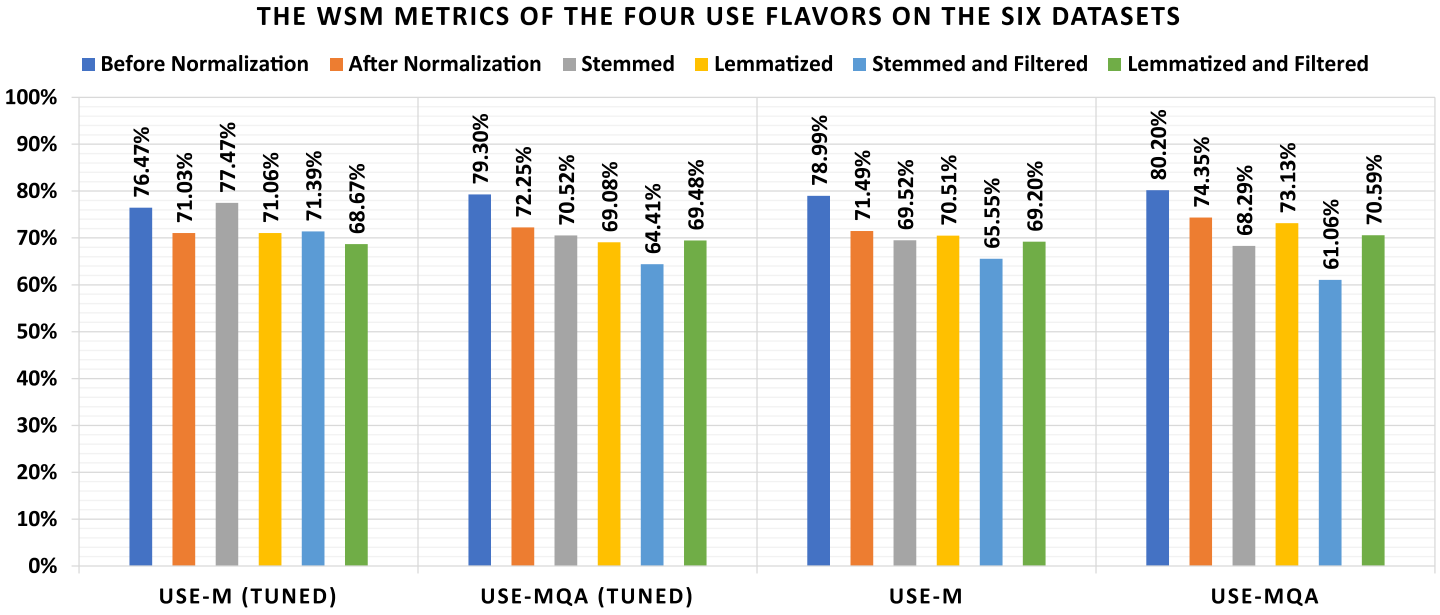


Figure 6 Graphical summarization of the WSM metrics of the four USE flavors on the six datasets.
Full-size DOI: 10.7717/peerj-cs.1070/fig-6

the highest WSM value of the different BERT models is (1) 95.26% using the "Before Normalization" dataset, (2) 94.67% using the "After Normalization" dataset, (3) 92.96% using the "Stemmed" dataset, (4) 94.40%, using the "Lemmatized" dataset, (5) 92.91% using the "Stemmed and Filtered" dataset, and (6) 93.49% using the "Lemmatized and Filtered" dataset. The best overall WSM value is 95.26%, produced by the "BERT-BV02T" model and the "Before Normalization" dataset. The lowest overall WSM value is 38.30%, produced by the "BERT-LV2" model and "Stemmed" dataset. The average WSM value is 79.33%. Table 14 summarizes the WSM metrics and presents the maximum values of each row and column. Also, they are summarized graphically in Fig. 7.

The BERT Experiments Remarks: It can be conduced from Tables 12–14, and Fig. 7 that (1) the "BERT-LV2" and "BERT-LV02T" are not performing well, (2) the lemmatization is better than stemming concerning all BERT models without filtering and

Table 12 The loss and confusion matrix results of the five BERT flavors on the six datasets.

Keyword	Dataset	Loss	TP	TN	FP	FN
BERT-BV2	Before Normalization	0.478	927	990	84	29
BERT-BV2	After Normalization	0.464	929	993	81	27
BERT-BV2	Stemmed	0.544	922	987	87	34
BERT-BV2	Lemmatized	0.443	922	1,008	66	34
BERT-BV2	Stemmed and Filtered	0.560	923	984	90	33
BERT-BV2	Lemmatized and Filtered	0.460	921	998	76	35
BERT-LV2	Before Normalization	0.692	0	1,074	0	956
BERT-LV2	After Normalization	0.691	0	1,074	0	956
BERT-LV2	Stemmed	0.664	24	1,050	24	932
BERT-LV2	Lemmatized	0.694	0	1,074	0	956
BERT-LV2	Stemmed and Filtered	0.692	0	1,074	0	956
BERT-LV2	Lemmatized and Filtered	0.371	904	1,005	69	52
AraElectra	Before Normalization	0.384	921	1,003	71	35
AraElectra	After Normalization	0.387	909	1,014	60	47
AraElectra	Stemmed	0.411	903	995	79	53
AraElectra	Lemmatized	0.403	898	1,006	68	58
AraElectra	Stemmed and Filtered	0.473	906	989	85	50
AraElectra	Lemmatized and Filtered	0.399	910	985	89	46
BERT-BV02T	Before Normalization	0.330	926	1,024	50	30
BERT-BV02T	After Normalization	0.409	931	1,005	69	25
BERT-BV02T	Stemmed	0.543	907	1,004	70	49
BERT-BV02T	Lemmatized	0.418	921	1,009	65	35
BERT-BV02T	Stemmed and Filtered	0.557	896	1,002	72	60
BERT-BV02T	Lemmatized and Filtered	0.495	919	992	82	37
BERT-LV02T	Before Normalization	0.692	0	1,074	0	956
BERT-LV02T	After Normalization	0.404	932	1,007	67	24
BERT-LV02T	Stemmed	0.691	0	1,074	0	956
BERT-LV02T	Lemmatized	0.468	923	1,012	62	33
BERT-LV02T	Stemmed and Filtered	0.538	908	1,002	72	48
BERT-LV02T	Lemmatized and Filtered	0.657	24	1,071	3	932

all BERT model with filtering unless the "BERT-LV02T" model, (3) the lemmatization without filtering is better than the lemmatization with filtering concerning all BERT models unless the "BERT-LV2" model, (4) the stemming without filtering is better than the stemming with filtering concerning all BERT models unless the "BERT-LV2" and "BERT-LV02T" models, and (5) there are no major differences between using the normalization or ignoring it.

Table 13 The performance metrics of the five BERT flavors on the six datasets.

Keyword	Dataset	Accuracy (%)	Balanced accuracy (%)	Precision (%)	Specificity (%)	Recall (%)	F1 (%)	IoU (%)	ROC (%)	Youden index (%)	NPV (%)	WSM (%)
BERT-BV2	Before Normalization	94.43	94.57	91.69	92.18	96.97	94.26	89.13	94.60	89.15	97.15	93.41
BERT-BV2	After Normalization	94.68	94.82	91.98	92.46	97.18	94.51	89.59	94.85	89.63	97.35	93.70
BERT-BV2	Stemmed	94.04	94.17	91.38	91.90	96.44	93.84	88.40	94.20	88.34	96.67	92.94
BERT-BV2	Lemmatized	95.07	95.15	93.32	93.85	96.44	94.86	90.22	95.16	90.30	96.74	94.11
BERT-BV2	Stemmed and Filtered	93.94	94.08	91.12	91.62	96.55	93.75	88.24	94.12	88.17	96.76	92.83
BERT-BV2	Lemmatized and Filtered	94.53	94.63	92.38	92.92	96.34	94.32	89.24	94.65	89.26	96.61	93.49
BERT-LV2	Before Normalization	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV2	After Normalization	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV2	Stemmed	52.91	50.14	50	97.77	2.51	4.78	2.45	69.15	0.28	52.98	38.30
BERT-LV2	Lemmatized	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV2	Stemmed and Filtered	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV2	Lemmatized and Filtered	94.04	94.07	92.91	93.58	94.56	93.73	88.20	94.07	88.14	95.08	92.84
AraElectra	Before Normalization	94.78	94.86	92.84	93.39	96.34	94.56	89.68	94.88	89.73	96.63	93.77
AraElectra	After Normalization	94.73	94.75	93.81	94.41	95.08	94.44	89.47	94.75	89.50	95.57	93.65
AraElectra	Stemmed	93.50	93.55	91.96	92.64	94.46	93.19	87.25	93.55	87.10	94.94	92.21
AraElectra	Lemmatized	93.79	93.80	92.96	93.67	93.93	93.44	87.70	93.80	87.60	94.55	92.52
AraElectra	Stemmed and Filtered	93.35	93.43	91.42	92.09	94.77	93.07	87.03	93.44	86.86	95.19	92.06
AraElectra	Lemmatized and Filtered	93.35	93.45	91.09	91.71	95.19	93.09	87.08	93.47	86.90	95.54	92.09
BERT-BV02T	Before Normalization	96.06	96.10	94.88	95.34	96.86	95.86	92.05	96.11	92.21	97.15	95.26
BERT-BV02T	After Normalization	95.37	95.48	93.10	93.58	97.38	95.19	90.83	95.50	90.96	97.57	94.50
BERT-BV02T	Stemmed	94.14	94.18	92.84	93.48	94.87	93.84	88.40	94.18	88.36	95.35	92.96
BERT-BV02T	Lemmatized	95.07	95.14	93.41	93.95	96.34	94.85	90.21	95.15	90.29	96.65	94.11
BERT-BV02T	Stemmed and Filtered	93.50	93.51	92.56	93.30	93.72	93.14	87.16	93.51	87.02	94.35	92.18
BERT-BV02T	Lemmatized and Filtered	94.14	94.25	91.81	92.36	96.13	93.92	88.54	94.27	88.49	96.40	93.03
BERT-LV02T	Before Normalization	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV02T	After Normalization	95.52	95.63	93.29	93.76	97.49	95.35	91.10	95.64	91.25	97.67	94.67
BERT-LV02T	Stemmed	52.91	50	N/A	100	0	N/A	0	70.71	0	52.91	40.82
BERT-LV02T	Lemmatized	95.32	95.39	93.71	94.23	96.55	95.11	90.67	95.39	90.78	96.84	94.40
BERT-LV02T	Stemmed and Filtered	94.09	94.14	92.65	93.30	94.98	93.80	88.33	94.14	88.28	95.43	92.91
BERT-LV02T	Lemmatized and Filtered	53.94	51.12	88.89	99.72	2.51	4.88	2.50	70.54	2.23	53.47	42.98

Table 14 The WSM metrics of the five BERT flavors on the six datasets.

	Before normalization (%)	After normalization (%)	Stemmed (%)	Lemmatized (%)	Stemmed and filtered (%)	Lemmatized and filtered (%)	MAX (%)
BERT-BV2	93.41	93.70	92.94	94.11	92.83	93.49	94.11
BERT-LV2	40.82	40.82	38.30	40.82	40.82	92.84	92.84
AraElectra	93.77	93.65	92.21	92.52	92.06	92.09	93.77
BERT-BV02T	95.26	94.50	92.96	94.11	92.18	93.03	95.26
BERT-LV02T	40.82	94.67	40.82	94.40	92.91	42.98	94.67
MAX	95.26	94.67	92.96	94.40	92.91	93.49	95.26

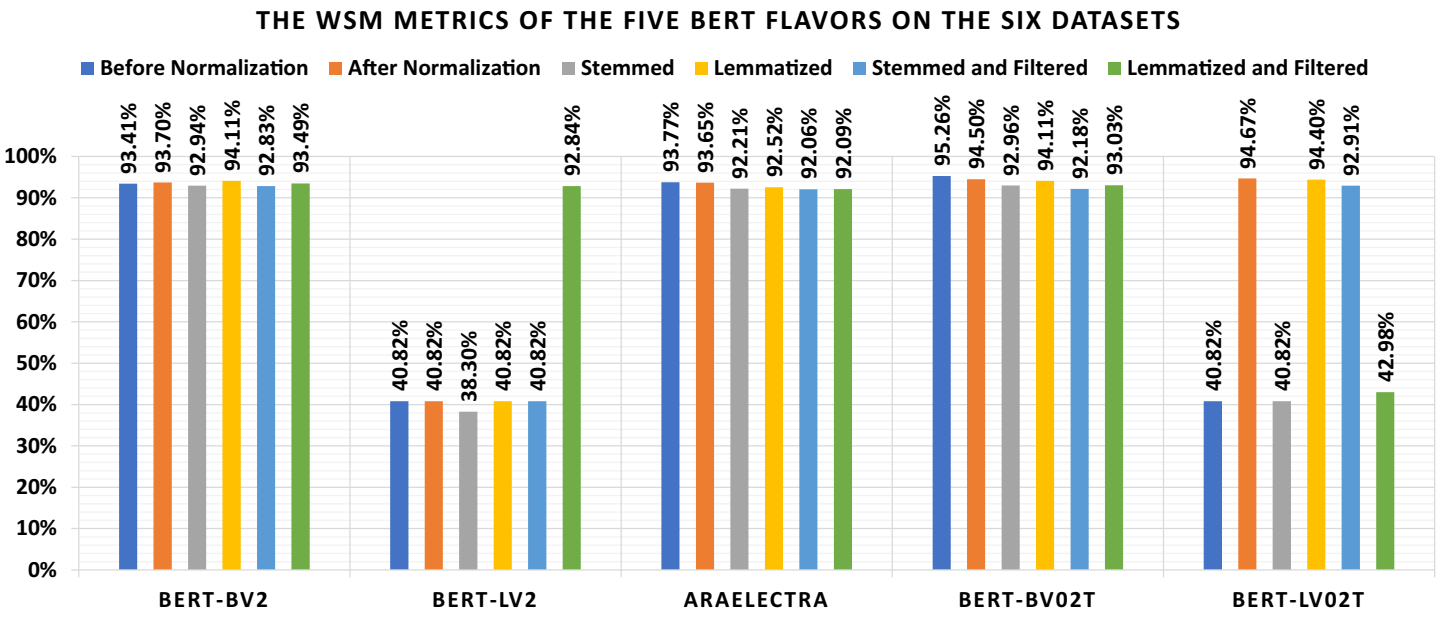


Figure 7 Graphical summarization of the WSM metrics of the five BERT flavors on the six datasets.Full-size DOI: 10.7717/peerj-cs.1070/fig-7

CONCLUSIONS

Various mental disorders such as MDD (or depression) affect millions of individuals worldwide. These disorders affect their daily lives. More than 300 million people worldwide are affected, and the number is growing daily. For example, it is estimated that 800K teens commit suicide yearly, making it the second leading cause of death in youth. Social networking sites such as Twitter and Facebook allow individuals to share their feelings and opinions on various subjects. The process also leads to the generation of social data, providing important insight into people's mental health and daily activities. In studying and analyzing user data, researchers can discover how mental health influences human behavior and how they can assist users in recovering. Researchers have used social media data to track depression and other mental illnesses. Social media data detect depressive symptoms discreetly before they progress to more severe stages. Consequently,

prevention and therapy strategies can be recommended to patients early. NLP is used to analyze spoken and written language. Languages like Arabic and English can be processed using NLP. SA identifies positive and negative aspects of a text, speech, or social media post. This method can analyze a person's depression level based on social media posts and negative sentiment ratings. The present study presents a complete framework for performing text classification (*i.e.*, tweets). The framework can be applied in six phases beginning with the data acquisition phase and transferring to classification through two cascading phases. The trained and optimized model classifies the Arabic tweets as "Suicide" or "Normal" using the trained and optimized model. The work also provides a preprocessing algorithm for Arabic tweets. Five different annotators have annotated a recently scraped Arabic tweets dataset, while the system performance was assessed using different performance metrics. General insights were obtained from the WSM. A USE model and a BERT model in Arabic were investigated. USE models have the best WSM at 80.20%, and Arabic BERT models have the best WSM at 95.26%. Different models can test a wider range of scraped data in future studies. The work can be extended by applying automatic learning for Arabic to accept and predict future tweets.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by Princess Nourah bint Abdulrahman University, Researchers Supporting Project number (PNURSP2022R293), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Princess Nourah bint Abdulrahman University, Researchers Supporting Project number: NURSP2022R293.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Nadiah A. Baghdadi analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Amer Malki performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Hossam Magdy Balaha conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Yousry AbdulAzeem conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

- Mahmoud Badawy conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Mostafa Elhosseini conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data is available at GitHub: https://github.com/HossamBalaha/An-Optimized-Deep-Learning-Approach-for-Suicide-Detection-through-Arabic-Tweets.

REFERENCES

Akhtar S, Ghosal D, Ekbal A, Bhattacharyya P, Kurohashi S. 2019. All-in-one: emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing* **13(1)**:285–297 DOI 10.1109/TAFFC.2019.2926724.

Alabdulkreem E. 2021. Prediction of depressed Arab women using their tweets. *Journal of Decision Systems* **30(2–3)**:102–117 DOI 10.1080/12460125.2020.1859745.

AlSagri HS, Ykhlef M. 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems* **103(8)**:1825–1832 DOI 10.1587/transinf.2020EDP7023.

Antoun W, Baly F, Hajj H. 2020. Arabert: transformer-based model for Arabic language understanding. In: *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*. 9.

Antoun W, Baly F, Hajj H. 2021. AraELECTRA: pre-training text discriminators for Arabic language understanding. In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, 191–195.

Asif M, Ishtiaq A, Ahmad H, Aljuaid H, Shah J. 2020. Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics* **48(3)**:101345 DOI 10.1016/j.tele.2020.101345.

Baghdadi NA, Malki A, Abdelaliem SF, Balaha HM, Badawy M, Elhosseini M. 2022a. An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based convolutional neural network. *Computers in Biology and Medicine* **144(1)**:105383 DOI 10.1016/j.compbiomed.2022.105383.

Baghdadi NA, Malki A, Balaha HM, Badawy M, Elhosseini M. 2022b. A3c-tl-gto: Alzheimer automatic accurate classification using transfer learning and artificial gorilla troops optimizer. *Sensors* **22(11)**:4250 DOI 10.3390/s22114250.

Bhushan B, Sharma N. 2021. Transaction privacy preservations for blockchain technology. In: *International Conference on Innovative Computing and Communications*. Cham: Springer, 377–393.

Chiu CY, Lane HY, Koh JL, Chen AL. 2021. Multimodal depression detection on Instagram considering time interval of posts. *Journal of Intelligent Information Systems* **56(1)**:25–47 DOI 10.1007/s10844-020-00599-5.

DBSA. 2022. Depression and bipolar support alliance. *Available at https://www.dbsalliance.org/* (accessed 16 April 2022).

Essien A, Petrounias I, Sampaio P, Sampaio S. 2021. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web* **24(4)**:1345–1368 DOI 10.1007/s11280-020-00800-3.

Google. 2019a. Universal-sentence-encoder-multilingual. *Available at https://tfhub.dev/google/universal-sentence-encoder-multilingual/3*.

Google. 2019b. Universal-sentence-encoder-multilingual-qa. *Available at https://tfhub.dev/google/universal-sentence-encoder-multilingual-qa/3*.

Islam M, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. 2018. Depression detection from social network data using machine learning techniques. *Health Information Science and Systems* **6(1)**:1–12 DOI 10.1007/s13755-018-0046-0.

Joshi A, Karimi S, Sparks R, Paris C, MacIntyre CR. 2019. A comparison of word-based and context-based representations for classification problems in health informatics. *ArXiv preprint* DOI 10.48550/arXiv.1906.05468.

Jyothi Prasanth D, Dhalia Sweetlin J, Sruthi S. 2022. Exploring human emotions for depression detection from twitter data by reducing misclassification rate. In: *Proceedings of International Conference on Computational Intelligence*. Cham: Springer, 127–135.

Kemp S. 2022. Digital 2022: global overview report. *Available at https://datareportal.com/reports/digital-2022-global-overview-report* (accessed 16 April 2022).

Kessler RC, Bromet EJ, de Jonge P, Shahly V, Wilcox M. 2017. The burden of depressive illness. Johns Hopkins University Press. *Available at https://psycnet.apa.org/record/2017-36588-003*.

Kim J, Lee J, Park E, Han J. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports* **10(1)**:1–6 DOI 10.1038/s41598-020-68764-y.

Kour H, Gupta MK. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications* **81(17)**:1–37 DOI 10.1007/s11042-022-12648-y.

Kroenke K, Spitzer RL, Williams JB. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16(9)**:606–613 DOI 10.1046/j.1525-1497.2001.016009606.x.

Lia RJ, Siddikk AB, Muntasir F, Rahman SSMM, Jahan N. 2022. Depression detection from social media using Twitter's tweet. In: *Big Data Intelligence for Smart Applications*. New York: Springer, 209–226.

Liu D, Feng XL, Ahmed F, Shahid M, Guo J. 2022a. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health* **9(3)**:e27244 DOI 10.2196/27244.

Liu Y, Luo X, Zhang M, Tao Z, Liu F. 2022b. Who are there: discover Twitter users and tweets for target area using mention relationship strength and local tweet ratio. *Journal of Network and Computer Applications* **199(FEB)**:103302 DOI 10.1016/j.jnca.2021.103302.

Mathers CD, Loncar D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLOS Medicine* **3(11)**:e442 DOI 10.1371/journal.pmed.0030442.

Mishra V, Garg T. 2018. A systematic study on predicting depression using text analytics. *Journal of Fundamental and Applied Sciences* **10(2)**:172123.

Nair SS, Ashok A, Divya Pai R, Hari Narayanan A. 2022. Detection of early depression signals using social media sentiment analysis on big data. In: *Computer Networks and Inventive Communication Technologies*. New York: Springer, 413–422.

NHS. 2019. Clinical depression—symptoms. *Available at https://www.nhs.uk/mental-health/conditions/clinical-depression/symptoms/* (accessed 16 April 2022).

Ni H, Wang S, Cheng P. 2021. A hybrid approach for stock trend prediction based on tweets embedding and historical prices. *World Wide Web* **24(3)**:849–868 DOI 10.1007/s11280-021-00880-9.

NLTK. 2022. Nltk: natural language toolkit. *Available at https://www.nltk.org/*.

Park J, Moon N. 2022. Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability* **14(6)**:3569 DOI 10.3390/su14063569.

Rao G, Zhang Y, Zhang L, Cong Q, Feng Z. 2020. MGL-CNN: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access* **8**:32395–32403 DOI 10.1109/ACCESS.2020.2973737.

Rosa RL, Schwartz GM, Ruggiero WV, Rodrguez DZ. 2018. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics* **15(4)**:2124–2135 DOI 10.1109/TII.2018.2867174.

Russell JA. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* **39(6)**:1161–1178 DOI 10.1037/h0077714.

Safa R, Bayat P, Moghtader L. 2022. Automatic detection of depression symptoms in Twitter using multimodal analysis. *The Journal of Supercomputing* **78(4)**:4709–4744 DOI 10.1007/s11227-021-04040-8.

Salas-Zárate R, Alor-Hernández G, Salas-Zárate MdP, Paredes-Valverde MA, Bustos-López M, Sánchez-Cervantes JL. 2022. Detecting depression signs on social media: a systematic literature review. In: *Healthcare*. Vol. 10. Basel: MDPI, 291.

Seliya N, Khoshgoftaar TM, Van Hulse J. 2009. A study on the relationships of classifier performance metrics. In: *2009 21st IEEE International Conference on Tools with Artificial Intelligence*. Piscataway: IEEE, 59–66.

Shankdhar A, Mishra R, Shukla N. 2022. An application of deep learning in identification of depression among twitter users. In: *International Conference on Innovative Computing and Communications*. Cham: Springer, 661–669.

Shetty NP, Muniyal B, Anand A, Kumar S, Prabhu S. 2020. Predicting depression using deep learning and ensemble algorithms on raw twitter data. *International Journal of Electrical and Computer Engineering* **10(4)**:3751 DOI 10.11591/ijece.v10i4.pp3751-3756.

Smys S, Raj JS. 2021. Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *Journal of Trends in Computer Science and Smart Technology (TCSST)* **3(1)**:24–39 DOI 10.36548/jtcsst.

Sood A, Hooda M, Dhir S, Bhatia M. 2018. An initiative to identify depression using sentiment analysis: a machine learning approach. *Indian Journal of Science and Technology* **11(4)**:1–6 DOI 10.17485/ijst/2018/v11i4/119594.

Suman SK, Shalu H, Agrawal LA, Agrawal A, Kadiwala J. 2020. A novel sentiment analysis engine for preliminary depression status estimation on social media. *ArXiv preprint* DOI 10.48550/arXiv.2011.14280.

Taghva K, Elkhoury R, Coombs J. 2005. Arabic stemming without a root dictionary. In: *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*. Vol. 1. Piscataway: IEEE, 152–157.

Tommasel A, Diaz-Pace A, Rodriguez JM, Godoy D. 2021. Capturing social media expressions during the COVID-19 pandemic in Argentina and forecasting mental health and emotions. *ArXiv preprint*. DOI 10.48550/arXiv.2101.04540.

Tong L, Liu Z, Jiang Z, Zhou F, Chen L, Lyu J, Zhang X, Zhang Q, Sadka A, Wang Y, Li L, Zhou H. 2022. Cost-sensitive boosting pruning trees for depression detection on twitter. *IEEE*

Transactions on Affective Computing. Epub ahead of print 25 January 2022 DOI 10.1109/TAFFC.2022.3145634.

TWINT. 2022. Twint—Twitter intelligence tool. *Available at https://github.com/twintproject/twint*.

Vanam H, Jeberson RRR. 2021. Analysis of Twitter data through big data based sentiment analysis approaches. *Materials Today: Proceedings* Epub ahead of print 4 January 2021 DOI 10.1016/j.matpr.2020.11.486.

Victor E, Aghajan ZM, Sewart AR, Christian R. 2019. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological Assessment* **31(8)**:1019–1027 DOI 10.1037/pas0000724.

Wang L, Niu J, Yu S. 2019. Sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* **32(10)**:2026–2039 DOI 10.1109/TKDE.2019.2913641.

World Health Organization (WHO). 2021. Depression. *Available at https://www.who.int/news-room/fact-sheets/detail/depression* (accessed 16 April 2022).

Wongkoblap A, Vadillo MA, Curcin V. 2022. Social media big data analysis for mental health research. In: *Mental Health in a Digital World*. Amsterdam: Elsevier, 109–143.

Zafar A, Chitnis S. 2020. Survey of depression detection using social networking sites via data mining. In: *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. Piscataway: IEEE, 88–93.

Zerrouki T. 2012. Qalsadi, Arabic mophological analyzer library for Python. *Available at https://qalsadi.readthedocs.io/en/latest/README.html*.

Zhang T, Schoene AM, Ji S, Ananiadou S. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine* **5(1)**:1–13 DOI 10.1038/s41746-022-00589-7.

Zogan H, Razzak I, Wang X, Jameel S, Xu G. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* **25(1)**:1–24 DOI 10.1007/s11280-021-00992-2.

PeerJ Computer Science

COVID-19 mass media coverage in English and public reactions: a West-East comparison *via* Facebook posts

Ahmad R. Pratama[1,*] and Firman M. Firmansyah[2,*]

[1] Department of Informatics, Universitas Islam Indonesia, Sleman, Daerah Istimewa Yogyakarta, Indonesia
[2] Department of Technology and Society, Stony Brook University, Stony Brook, New York, United States
* These authors contributed equally to this work.

ABSTRACT

Newspapers and other mass media outlets are critical in shaping public opinion on a variety of contemporary issues, including the COVID-19 pandemic. This study examines how the pandemic is portrayed in the news and how the public reacted differently in the West and East using archival data from Facebook posts about COVID-19 news by English-language mass media between January 2020 and April 2022 (N = 711,646). Specifically, we employed the Valence Aware Dictionary and sEntiment Reasoner (Vader) to measure the news tone on each COVID-19 news item shared on Facebook by mass media outlets. In addition, we calculated a polarity score based on Facebook special reactions (*i.e.*, love, angry, sad, wow, haha, and care) received by each post to measure public reactions toward it. We discovered that people in Western countries reacted significantly more negatively to COVID-19 news than their East counterparts, despite the fact that the news itself, in aggregate, generally contained a relatively similar level of neutral tone in both West and East media. The implications of these distinctions are discussed in greater detail.

Subjects Data Science, Natural Language and Speech, Network Science and Online Social Networks, Text Mining, Sentiment Analysis
Keywords Facebook posts, Public reactions, COVID-19 news, Mass media, Sentiment analysis, West-East, Individualism, Collectivism, News tone

Submitted 30 May 2022
Accepted 30 August 2022
Published 28 September 2022

Corresponding authors
Ahmad R. Pratama,
ahmad.rafie@uii.ac.id
Firman M. Firmansyah,
manda.firmansyah@stonybrook.edu

Academic editor
Arkaitz Zubiaga

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.1111**

Copyright
2022 Pratama and Firmansyah

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Newspapers and other forms of mass media play a crucial role in shaping public opinion on a variety of contemporary issues, including the COVID-19 pandemic (*Medina, Rodriguez & Sarmiento, 2021*; *Rozado, 2021*; *Wu, 2021*). Even more so, they not only shaped how people perceive the outbreak (*Motta, Stecula & Farhart, 2020*), but also influenced how people cope with the disease and perform preventive measures against it. For example, in the United States, greater mass media consumption has been associated with a better social distancing practice (*Pedersen & Favero, 2020*) and higher COVID-19 vaccine acceptance rate (*Piltch-Loeb et al., 2021*). Meanwhile in Pakistan, mass media has been one of the reasons why people are otherwise reluctant to get vaccinated (*Ahmad Kamboh, Ittefaq & Sahi, 2022*; *Chaudhary et al., 2021*). In this regard, their journalism styles that commonly employ striking headlines with visual appeal arguably have nudged

Pakistani readers to generalize several cases of reported vaccine adverse side-effects, which subsequently make them against it (*Chaudhary et al., 2021*).

As the contrasting examples above suggest, it appears that the way in which the COVID-19 pandemic is portrayed in the news determines whether the media help or hinder the mitigation efforts. Indeed, even in the United States, where mass media in general helped prevent the spread of the virus as previously mentioned (*Pedersen & Favero, 2020*; *Piltch-Loeb et al., 2021*), some right-leaning media indirectly contributed to its spread by disseminating or even amplifying COVID-19 misinformation (*Motta, Stecula & Farhart, 2020*). Further analysis confirmed that their frequent readers tended to hold false beliefs about the pandemics and to disregard preventive measures such as social distancing (*Motta, Stecula & Farhart, 2020*; *Pedersen & Favero, 2020*; *Zhao et al., 2020*).

Considering the double-edged sword effect mass media can generate, it is important to examine how different media portray the outbreak differently. While past studies have initiated such efforts, they mostly focused on either certain news outlets *e.g.*, (*Hossain, Wahab & Khan, 2022*; *Ogbodo et al., 2020*), specific locations sharing similar cultural backgrounds *e.g.*, (*Mach et al., 2021*; *Rianto & Pratama, 2021*), or relatively small number of different countries (*Wu, 2021*). It is arguably difficult to infer a larger direct comparison, such as the differences between West and East mass media portrayals of the pandemic, let alone examine public reactions to such portrayals, due to these scopes. This study aims to fill this gap in the literature. Specifically, we intend to address the following research questions:

1. How have the mass media covered the COVID-19 pandemic in the past 2 years? Were there any differences in the tone used by mass media in the West compared to the East in their COVID-19 news coverage?
2. Did the audience in the West and East react differently toward COVID-19 pandemic news shared by mass media outlets?
3. Is there any relationship between the COVID-19 news tone by mass media and the public reactions it received?

To answer those questions, we analyzed archival data of Facebook posts by mass media in 18 countries between February 15, 2020, and April 14, 2022, along with their public reactions. Prior to getting into further details, let us highlight some differences between West and East societies in responding to the pandemic that can be explained through the individualism-collectivism cultural dimension (*Hofstede, Hofstede & Minkov, 2005*) along with socio-economic-political contexts and why Facebook posts and special reactions can help us uncover the sentiment behind them in the following sections.

COVID-19 and individualism-collectivism cultural framework

Individualism-collectivism framework is among the cultural characteristics that distinguish Western civilizations from their Eastern counterparts (*Hofstede, Hofstede & Minkov, 2005*). Individualism culture is characterized by a desire for a loosely knit social structure in which people are expected to care exclusively for themselves and their

immediate families (*Hofstede, 2011*). In contrast, collectivism expresses a preference for a tightly knit social structure in which individuals may expect their family or members of a certain ingroup to care for them in exchange for unquestioned allegiance (*Hofstede, 2011*). In simpler words, individualism culture emphasizes "I" and "me" while collectivism culture emphasizes "we" and "us". While these distinctions hold true at the group level, it may not always be the case at the individual level (*Hofstede, Hofstede & Minkov, 2005*). In this respect, individuals of the same nation, culture, or civilization may have different levels of individualism/collectivism themselves (*Hofstede & McCrae, 2004*), which are then referred to as individualism/collectivism personality traits (*Burton et al., 2021*; *Triandis, 2001*).

Recent studies revealed that the individualism-collectivism cultural differences also manifested in the different responses between people in Western and Eastern countries toward the COVID-19 mitigation policies and prevention efforts. For instance, *Chen, Frey & Presidente (2021)* analyzed daily data on geographic mobility in 111 countries during the early epidemic and discovered that people were less likely to adhere to lockdown regulations in countries with a higher degree of individualism such as the United States, compared to countries with a higher degree of collectivism such as China. Furthermore, *Leonhardt & Pezzuti (2022)* surveyed 400,000 people in more than 50 countries and found that people from collectivistic countries tended to have positive attitudes towards COVID-19 vaccine compared to those from individualistic countries. Connecting the dots, it is less surprising that countries with a greater degree of individualism had higher numbers of confirmed cases and mortality rates than countries with a greater degree of collectivism (*Chen & Biswas, 2022*; *Maaravi et al., 2021*; *Rajkumar, 2021*; *Ritchie et al., 2020*).

COVID-19 and socio-economic-political factors

Indeed, such a cultural framework is insufficient to explain why some countries in certain regions have higher COVID-19 confirmed cases and mortality rates than others. Other aspects such as socio-economic-political factors seem also to play roles. In terms of confirmed cases, for instance, the rate in a given country has been arguably contingent on the number of available testing kits. In Western countries such as the United States, COVID-19 tests were not only more available but also free (*Assistant Secretary for Health, 2020*). This condition helped authorities track and record such cases. Meanwhile, in Eastern countries such as Indonesia, people had to pay out of pocket to get tested (*Massola, 2020*). This condition arguably not only hindered Indonesians from getting tested but also made it harder to estimate the real number of positive cases in the country. It is not impossible that Indonesia had the same rate of positive cases as the United States, or even greater. Yet, it is difficult to confirm since the testing tools were inadequate in the first place.

The scarcity not only applies to COVID-19 testing kits but also to COVID-19 vaccines. Interestingly, when countries in both regions have arguably similar resources, vaccination rates in the Eastern countries such as Singapore tended to be higher than those in Western countries such as the United Kingdom (*Ritchie et al., 2020*). While it can be attributed to the cultural framework as previously mentioned, *i.e.*, (*Leonhardt & Pezzuti, 2022*), it could

also be due to different political situations. In the Western countries which are predominantly democratic, the governments have less authority to force people to get vaccinated. In contrast, in the Eastern countries, in which the governments are arguably more authoritative despite some of them being democratic (*e.g.*, Singapore), it is not difficult to impose such rules that forcely nudge their citizens to get vaccinated (*Dyer, 2021*).

Sentiment analysis through facebook posts and special reactions

In comparison to traditional approaches, big data and social media footprints make it easier to undertake nationwide or worldwide sentiment research. In addition, digital footprints on social media capture things as they have occurred, which is not always the case in traditional approaches like surveys and interviews. As a result, these digital footprints have also been extensively employed to study a wide range of social topics on the internet and in particular on social media over the past decade, including those that are quite sensitive such as anonymous online donations (*Firmansyah & Pratama, 2021*), racial digital identities (*Firmansyah & Jones, 2019*), and political preferences (*Oliveira, Bermejo & dos Santos, 2017*).

Apart from being, still, the largest social media in the world as of 2022 (*Statista, 2022*), Facebook also offers a unique way of conducting sentiment analysis studies using its special reactions, as illustrated in Fig. 1. While the reactions themselves in general are intriguing to study (*e.g.*, females, older, less educated users tend to give more 'likes' – *Hong, Chen & Li, 2017*), the special reactions (*i.e.*, love, wow, haha, sad, angry, and care) which were introduced in 2016, have been used to complement text-based sentiment analysis by researchers all around the globe (*Freeman, Alhoori & Shahzad, 2020*; *Oueslati, Khalil & Ounelli, 2018*; *Pratama, 2022*; *Tian et al., 2017*). As a result, it is possible to analyze both the content of the Facebook posts using text-based tools like the Valence Aware Dictionary for sEntiment Reasoning (Vader) (*Hutto & Gilbert, 2014*) as well as how the public reacted to those posts based on the special reactions they received.

MATERIALS AND METHODS

Data source

We used CrowdTangle (https://www.crowdtangle.com/), a free public insight tool owned and operated by Facebook to collect our dataset in this study (*CrowdTangle Team, 2019*). This tool allows verified researchers to query databases of more than seven million publicly available Facebook pages, groups, and profiles and their public posts. It is worth highlighting that Meta, the parent company of Facebook, does not allow researchers to perform any data scraping on any of their products even if the researchers aim for a greater good and academic purposes (*Clark, 2021*; *Vittorio, 2021*).

Data collection procedures

We used the search feature on CrowdTangle using the following parameters: Keyword = covid-19; Platform = Facebook; Account Type = Pages; Timeframe = March 1, 2020–February 28, 2022; Post Type = Links; Language = English, List = Asia Media, Australia

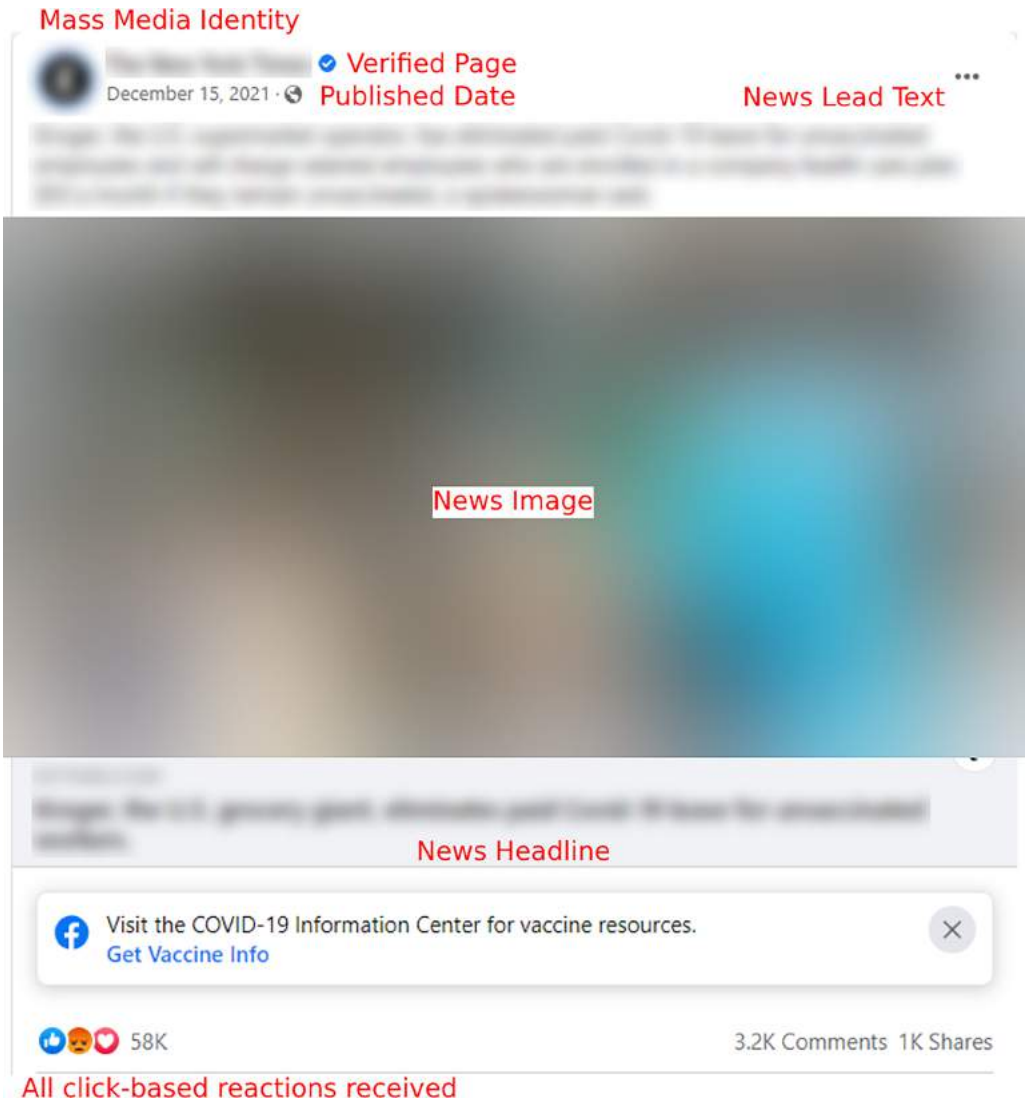


Figure 1 **An example of a news article about the COVID-19 pandemic along with special reactions it received.** Full-size DOI: 10.7717/peerj-cs.1111/fig-1

Media, Canada Media, UK Media, and US General Media (all managed and maintained by the CrowdTangle team) while keeping any other default settings. This query resulted in a total of 1,003,332 Facebook posts that received a total of 506,401,516 interactions by Facebook users. Each post is essentially a news article shared on Facebook by the respective mass media that links it to the complete news article on their own websites.

In the raw dataset, there were 25 countries in total based on the "Page Admin Top Country" column, some of which had only a small number of posts within the timeframe. Upon further inspection, we decided to omit several countries either due to very limited number of posts such as Denmark (4), South Korea (3), and Taiwan (3) or because their geographical location is not a clear cut to fall within the West or East category, namely Botswana, Israel, Qatar, and Russia. As a result, our final dataset, as summarized in

Table 1 Number of mass media's Facebook pages, posts, and their public reactions by region, subregion, and country. Individualism scores were retrieved from *Hofstede, Hofstede & Minkov (2005)*, which did not include scores for all countries.

Region	Subregion	Country	Number of pages	Number of posts	Total interactions	Total special reactions	Individualism score
West	Europe	Ireland	11	2,951	234,932	35,678	70
		Isle of Man	1	543	39,862	10,548	NA
		United Kingdom	327	109,773	47,447,149	12,032,252	89
		Subtotal	339	113,267	47,721,943	12,078,478	*Mean* = 79.5
	North America	Canada	62	208,459	62,402,186	16,461,408	80
		United States	69	114,417	239,334,047	81,016,163	91
		Subtotal	131	322,875	301,736,233	97,477,571	*Mean* = 85.5
	Oceania	Australia	41	50,686	23,995,733	6,911,059	90
	Total		511	486,828	373,453,909	116,467,108	*Mean* = 84.0
East	East Asia	China	2	1,201	1,446,111	38,548	20
	Southeast Asia	Cambodia	1	91	10,666	593	NA
		Indonesia	2	76	5,480	1,168	14
		Myanmar	3	1,345	260,852	7,350	NA
		Philippines	20	95,559	71,455,137	22,068,064	32
		Singapore	14	46,674	29,045,912	6,099,392	20
		Thailand	16	10,221	1,189,140	241,527	20
		Subtotal	46	153,966	101,967,187	28,418,094	*Mean* = 21.5
	South Asia	Bangladesh	4	69	62,305	4,441	20
		India	41	67,709	22,091,258	3,721,538	48
		Nepal	5	1,526	140,720	20,402	30
		Pakistan	6	304	108,077	8,583	14
		Sri Lanka	3	43	27,842	3,175	35
		Subtotal	59	69,651	22,430,202	3,758,139	*Mean* = 29.4
	Total		107	224,818	125,843,500	32,214,781	*Mean* = 25.3

Table 1, consists of 711,646 Facebook posts from 618 mass media in 18 countries, receiving a total of 499,297,409 interactions, 148,681,889 of which were special reactions.

As supplementary information, we added individualism scores for each country from Hofstede's dataset (*Hofstede, Hofstede & Minkov, 2005*) in Table 1. The individualism score ranges from 0 to 100, with higher values indicating higher individualism and in turn lower collectivism, and vice versa. As can be seen, countries in the West have relatively higher scores of individualism (M = 84.0, SD = 8.97) compared to countries in the East (M = 25.3, SD = 10.77). We also added cumulative confirmed COVID-19 cases and deaths per million people as well as vaccination rate for each country included in this study as presented in Figs. 2–4 respectively (*Ritchie et al., 2020*). It is worth noting that the terms "West" and "East" in this paper refer to civilizations and cultures rather than geographical locations. Thus, even though Australia is located in the east longitude and the southern hemisphere, it is still considered as part of the West.

Data analysis

We analyzed the data in three stages. The first stage serves as a preliminary which anchors further analyses. The second stage aims to answer the first two research questions. The

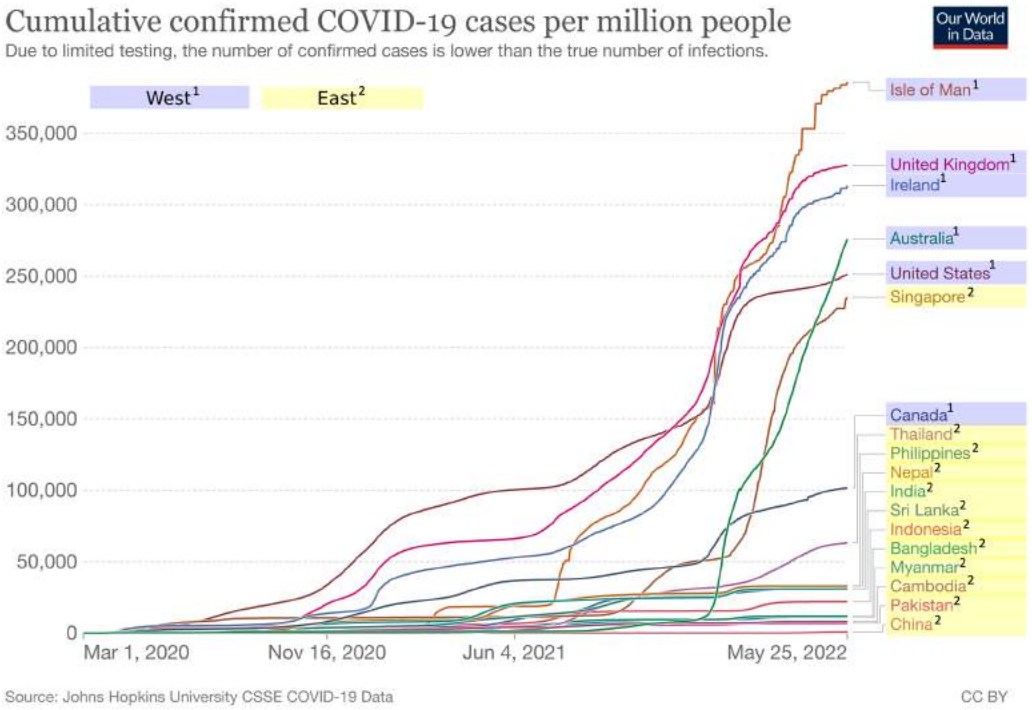


Figure 2 Cumulative confirmed COVID-19 cases per million people between West and East. Each line represents a country in either West (blue background) or East (yellow background). Figure source: Our World in Data, CC-BY 4.0, https://ourworldindata.org/explorers/coronavirus-data-explorer.
Full-size DOI: 10.7717/peerj-cs.1111/fig-2

third stage aims to answer the last research question. All the analyses were conducted using Python on Google Colab. The codes are available on the authors' GitHub repository (https://github.com/ahmadrafie/fbcovid19newssentiment).

First stage

In the first stage, we applied a text analysis counting the frequency of words in the news then visualizing the results using word clouds with the help of the "wordcloud" package (*Oesper et al., 2011*). In addition to the default stop words (*e.g.*, "and", "or", "only") and punctuations (*e.g.*, ".", ";", "?"), we excluded some other words (*e.g.*, "covid", "covid-19", "coronavirus") and country/city names (*e.g.*, "India", "Canada", "Manila") from the analysis. In this stage, we identified four relevant keywords that were used to subset the dataset for the subsequent specific-keyword analysis. As the subsetting process was mainly based on the identified keywords used in the news, the same news article could be analyzed more than once with different keywords. The subsetting process also made it possible for some articles to be included only in the general analysis (using all dataset) while being excluded from every single specific-keyword analysis (datasets with specific keywords). In other words, the subsets are not mutually exclusive for all news articles.

Second stage

In the second stage, we treated our dataset as cross-sectional data. We evaluated the content of the news shared by mass media before making comparisons between the West

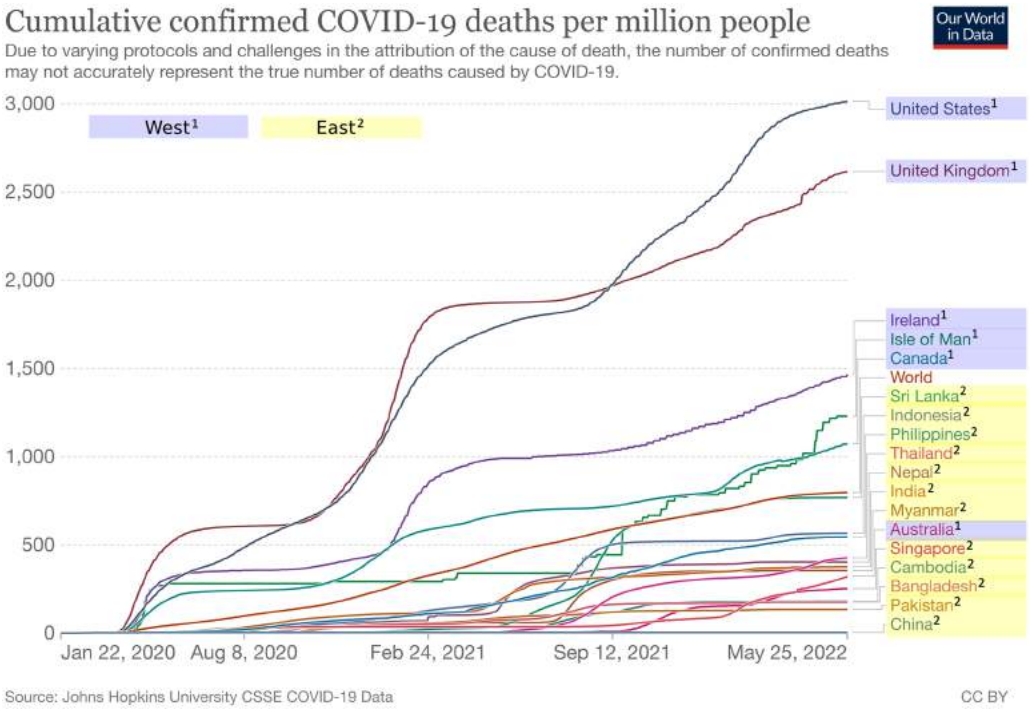


Figure 3 Cumulative confirmed COVID-19 deaths per million people between West and East. Each line represents a country in either West (blue background) or East (yellow background). Figure source: Our World in Data, CC-BY 4.0, https://ourworldindata.org/explorers/coronavirus-data-explorer.
Full-size DOI: 10.7717/peerj-cs.1111/fig-3

and East, both in the tone used in the news and the public reactions it received. We utilized the "vaderSentiment" package on Python (*Hutto & Gilbert, 2014*) to measure the news tone. More specifically, we used the Vader compound score, ranging from −1 for the most negative tone to 1 for the most positive tone for each Facebook post.

As for public reactions, we calculated a polarity score based on special reactions a Facebook post received using the formula specified in the previous studies in the literature (*Freeman, Alhoori & Shahzad, 2020*; *Pratama, 2022*). In doing so, we first utilized the 'love' reaction to indicate a positive valence (1) as opposed to 'sad' and 'anger' reactions for a negative valence (−1). Then we calculated an intensity score, which is a ratio of all special reactions received by a post to all click-based reactions (including "Like"). Finally, a polarity score is computed by multiplying the valence score by the intensity score. Similar to the Vader compound score, the polarity scores also range from −1 for the most negative reactions to 1 for the most positive reactions.

Once we obtained both Vader compound and polarity scores for each Facebook post in our dataset, we then performed a series of Welch's t-tests to examine the differences between news tone and public reactions in the West *vs.* in the East. The Welch's t-test is more appropriate for our dataset given the unequal variances and unequal sample size between the two categories (*Delacre, Lakens & Leys, 2017*). We also computed Cohen's d effect size (*Fritz, Morris & Richler, 2012*) to help interpret the t-test results. We repeated the analysis using smaller subsets of the data by using the top four most relevant keywords

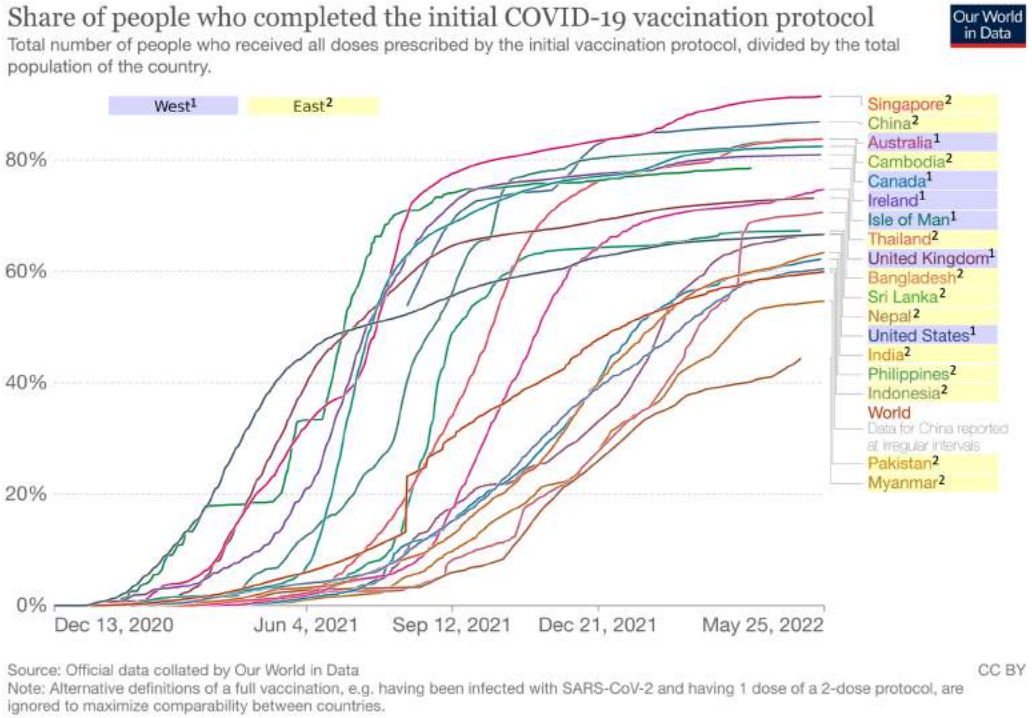


Figure 4 COVID-19 vaccination rate in West and East. Each line represents a country in either West (blue background) or East (yellow background). Figure source: Our World in Data, CC-BY 4.0, https://ourworldindata.org/explorers/coronavirus-data-explorer. Full-size DOI: 10.7717/peerj-cs.1111/fig-4

in the text to give a more detailed understanding of the differences. This process was done with the help of "Researchpy" package (*Bryant, 2022*).

Third stage

In the third stage, we aggregated the data by month and region. We visualized the data using line charts and conducted a correlation test between the news tone and the public reactions for both the West and East media. Similar to the second stage, we repeated the analysis using the same smaller subsets of data where we filtered our dataset by four specific-keywords. This process was also done with the help of "Researchpy" package (*Bryant, 2022*).

RESULTS

COVID-19 news coverage in the West and East media

Figure 5 shows the word cloud visualizations of the news article posted on Facebook by mass media, either overall or broken down by regions and sentiment of the posts. As previously mentioned, stop words and punctuations were excluded from the preliminary text analysis. Here, we found four relevant keywords, *i.e.*, "case", "vaccine", "health", and "death" for the further specific-keyword analysis to supplement the general analysis based on the overall dataset. It should be noted, we did not consider some other words like "pandemic", "new", or "say" relevant since the former is too similar to the main topic of the

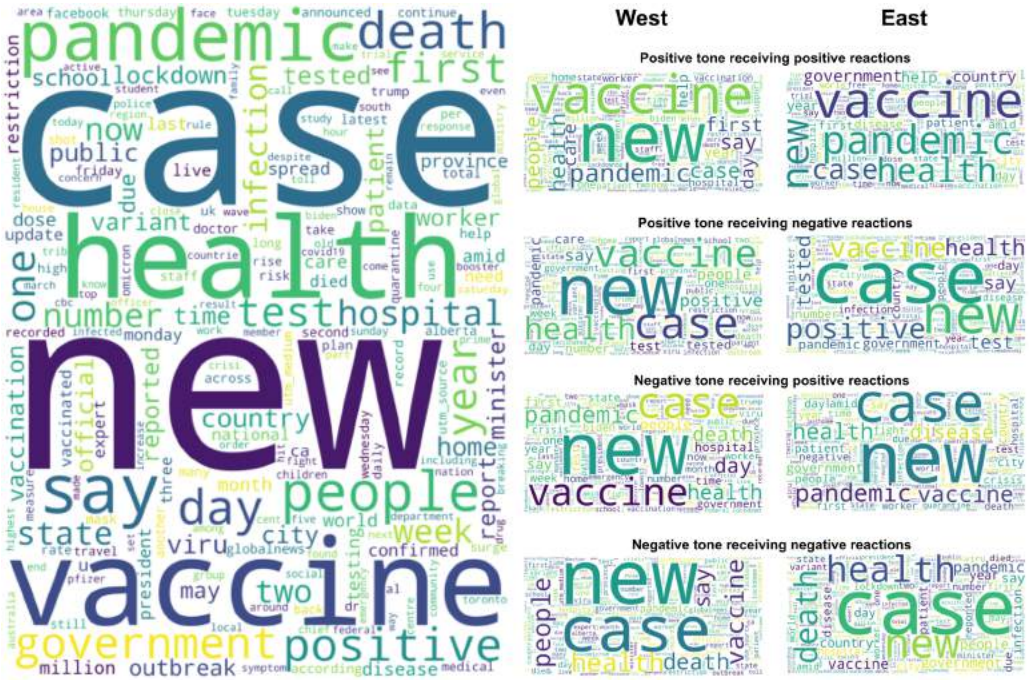


Figure 5 **Word cloud visualizations of COVID-19 news posted on Facebook by mass media.**
Full-size DOI: 10.7717/peerj-cs.1111/fig-5

overall dataset (*i.e.*, "COVID-19") while the other two are not a noun, making them too general and thus arguably make less sense to be analyzed separately.

Figure 5 (right side) further shows which relevant keywords appear more frequently in which news tone and which public reactions. For example, the word "case" received mostly negative reactions in both regions, but positively toned news containing this keyword is more prevalent in the East than in the West. Another example, the word "vaccine" is mostly covered in a positive tone, receiving more positive than negative public reactions in both regions, but negatively toned news containing this keyword is more common in the West than in the East.

COVID-19 news tone and public reactions in the West and East

Table 2 shows the percentages of positive and negative public reactions to two different news tones: positive and negative in both regions. As can be seen, in the West region, regardless of the news tone, in general, the public reacted more negatively. This finding applies to all subregions, including Europe, North America, and Oceania. Meanwhile, in the East region, the public reactions show more variations. In aggregate, they reacted more positively only toward positive news tones and more negatively toward negative tones. Interestingly, in the East Asia subregion, the public reacted more positively to all news regardless of its tonality. Even more so, the percentage of those reacting negatively is less than half of those reacting positively.

Furthermore, Table 3 presents the calculated Vader compound scores for the news tone and polarity scores for the public reactions in both the West and East regions, along with

Table 2 Percentage of positive and negative news tone in mass media's Facebook posts by region, subregion, and public reactions.

Region	Subregion	Public reactions	News tone	
			Positive	Negative
West	Europe	Positive	38.39%	21.70%
		Negative	61.61%	78.30%
	North America	Positive	35.50%	22.52%
		Negative	64.50%	77.48%
	Oceania	Positive	31.29%	21.97%
		Negative	68.71%	78.03%
		Positive	35.72%	22.26%
		Negative	64.28%	77.74%
East	East Asia	Positive	82.81%	73.08%
		Negative	17.19%	26.92%
	Southeast Asia	Positive	50.72%	33.11%
		Negative	49.28%	66.89%
	South Asia	Positive	58.59%	42.72%
		Negative	41.41%	57.28%
		Positive	53.46%	36.37%
		Negative	46.54%	63.63%

Table 3 **Welch's t-test results comparing news tone and public reactions in the West and East regions for the entire dataset and all four subsets.** Numbers reported in the West and East columns are means with standard deviations within parentheses.

Dataset	Variable	West	East	*p*-value	Cohen's d	Effect size
All	News Tone	−0.003 (0.579)	0.034 (0.605)	<0.001	0.063	Negligible
	Public Reactions	−0.317 (0.418)	−0.106 (0.333)	<0.001	0.521	Medium
Case	News Tone	0.022 (0.511)	−0.011 (0.546)	<0.001	0.062	Negligible
	Public Reactions	−0.445 (0.392)	−0.203 (0.289)	<0.001	0.668	Medium
Vaccine	News Tone	0.096 (0.546)	0.148 (0.553)	<0.001	0.094	Negligible
	Public Reactions	−0.202 (0.432)	−0.010 (0.346)	<0.001	0.471	Medium
Health	News Tone	0.005 (0.569)	−0.014 (0.588)	<0.001	0.034	Negligible
	Public Reactions	−0.354 (0.431)	−0.154 (0.331)	<0.001	0.498	Medium
Death	News Tone	−0.263 (0.555)	−0.374 (0.556)	<0.001	0.200	Small
	Public Reactions	−0.531 (0.368)	−0.244 (0.271)	<0.001	0.827	Large

Welch's t-test results. It should be noted, different from Table 2, Table 3 also presents the results for subsetted datasets based on relevant keywords as previously explained. As can be seen, while all the t-test results indicate significant differences in the news tone between West and East media, only "death" related news has a non-negligible effect size, thus meaningful. In this regard, the mass media in the East portrayed death-related news with a slightly more negative tone than the mass media in the West did. Meanwhile, for all general COVID-19 news, case-related news, vaccine-related news, and health-related news, both West and East media, in aggregate, had a relatively similar neutral tone.

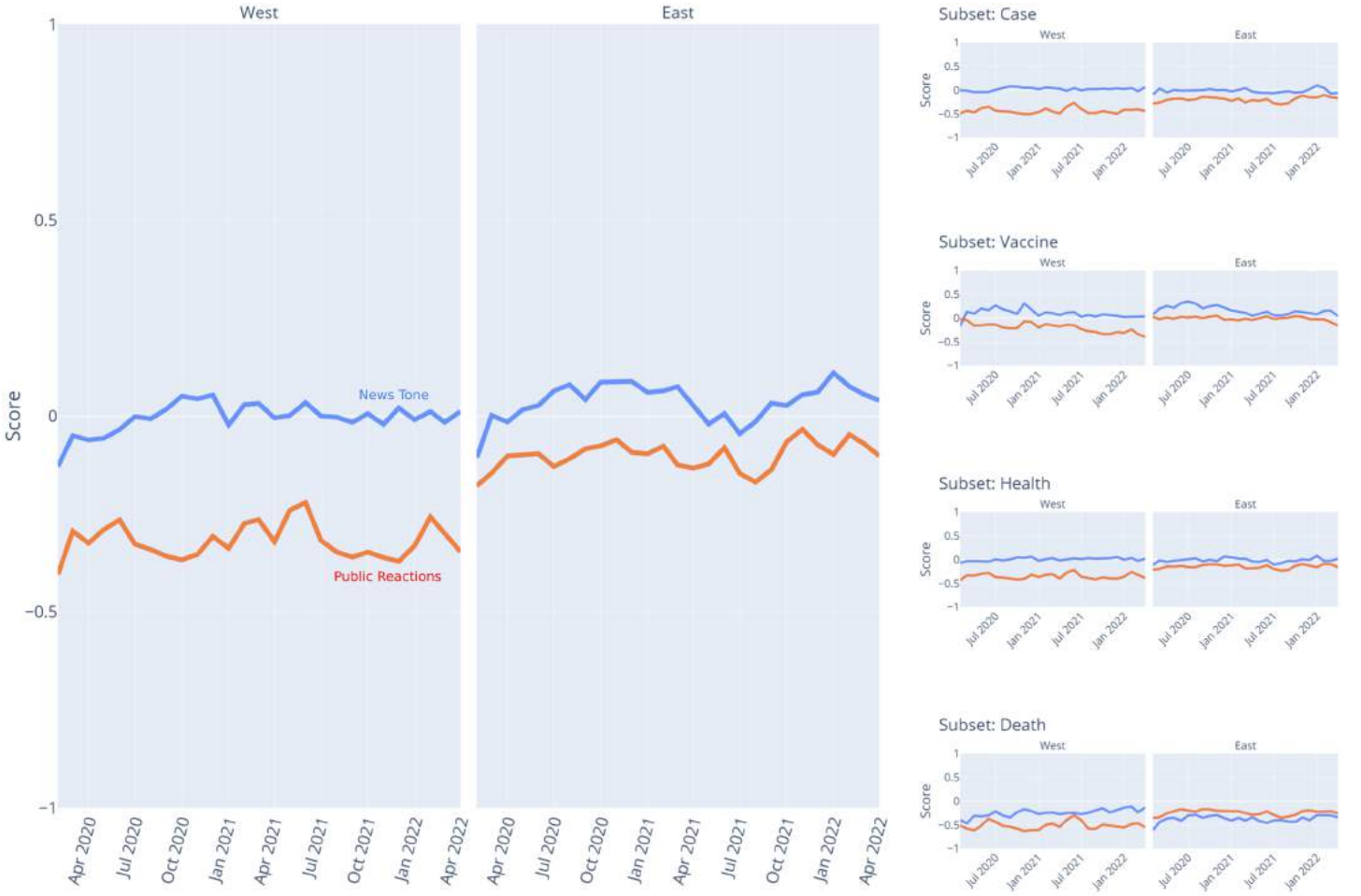


Figure 6 COVID-19 news tone and public reactions on Facebook over months. The blue line represents news tone as measured by the Vader compound score, while the orange line represents public reactions as measured by the polarity score based on special reactions.
Full-size DOI: 10.7717/peerj-cs.1111/fig-6

As for the public reactions, our analysis results indicate that the overall COVID-19 news in the West significantly received more negative reactions than COVID-19 news in the East by the public. Interestingly, these meaningful differences also apply to all specific keywords (*i.e.*, "case", "vaccine", "health", "death") as indicated by medium to large Cohen's d effect sizes.

Moreover, Figs. 6 and 7 present the time series analysis results for the Vader compound scores and public reactions' polarity scores for the general dataset and specific related keywords as well as by subregions, respectively. The visualizations highlight the fact that, aside from the previously mentioned death-related news and the East Asia subregion, the news tones were not significantly different between the mass media in the West and the East. On the other hand, when it comes to public reactions, the COVID-19 news in the West received more negative public reactions than similar news in the East, respectively.

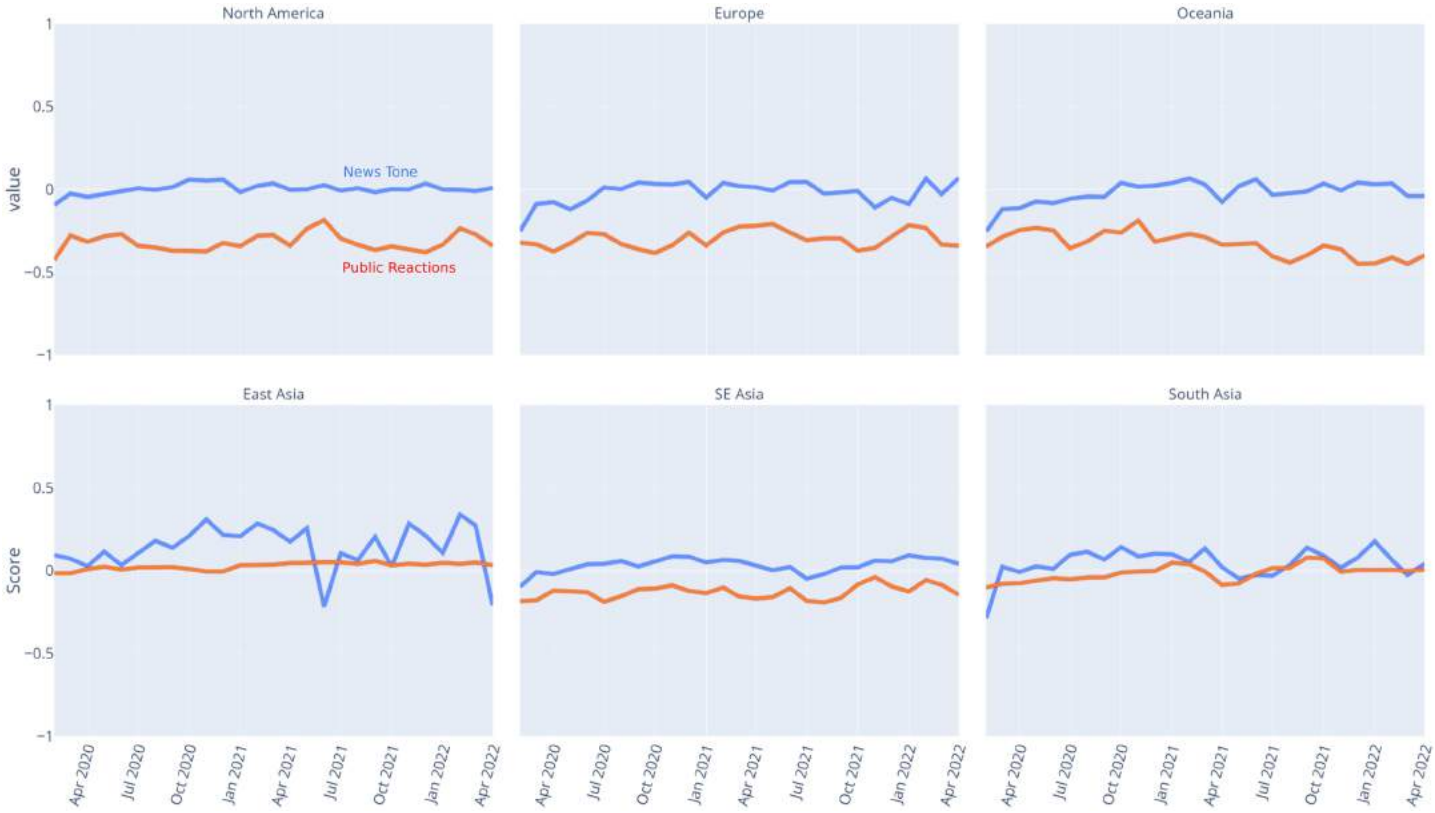


Figure 7 COVID-19 news tone and public reactions on Facebook over months across subregions. The blue line represents news tone as measured by the Vader compound score, while the orange line represents public reactions as measured by the polarity score based on special reactions.
Full-size DOI: 10.7717/peerj-cs.1111/fig-7

Table 4 Correlation test results between news tone and public reactions (monthly aggregate) in the West and East regions for the entire dataset and all four subsets.

Dataset	Region	r value	*p*-value	Correlation strength
All	West	0.227	0.255	Not significant
	East	0.626	<0.001	Strong positive
Case	West	−0.291	0.142	Not significant
	East	0.304	0.123	Not significant
Vaccine	West	0.313	0.112	Not significant
	East	0.431	0.025	Moderate positive
Health	West	−0.004	0.983	Not significant
	East	0.376	0.053	Not significant
Death	West	−0.014	0.943	Not significant
	East	0.596	0.001	Moderate positive

Relationships between COVID-19 news tone and public reactions

Table 4 summarizes the correlations between news tones and public reactions in the monthly aggregate. As shown, in general, there is no association between news tones and public reactions in the West. Public reactions in the West are consistently negative

regardless of the tone used by the mass media in COVID-19 news. In contrast, in the East, news tones strongly correlate with public reaction for all datasets in a positive direction. They were also moderately correlated in a positive direction for "vaccine" and "death" related news. In other words, the more positive the tone used by the mass media in the East in news related to the COVID-19 pandemic, including vaccine and death, the more positive the public reacts. Likewise, the more negative the tone used by the mass media in the East in news, the more negative the public reacts.

DISCUSSION

The analysis results using Vader (*Hutto & Gilbert, 2014*) have revealed that in aggregate, all news articles in English related to the COVID-19 pandemic shared on Facebook by mass media in both West and East regions had relatively similar levels of neutral tone. To be precise, they differed in very minor ways: the East media had a slightly more positive tone (*i.e.*, $M = 0.034$, SD = 0.605) while the West media had a slightly negative tone (*i.e.*, $M = -0.003$, SD = 0.579). However, the fact that their difference, albeit statistically significant, has negligible effect size suggests that there is no meaningful difference between them. This significance is mostly driven by the very large sample size in our dataset.

Furthermore, even though some mass media in both West and East regions seemed to intentionally frame the pandemic more negatively as the literature suggests (*Ahmad Kamboh, Ittefaq & Sahi, 2022*; *Chaudhary et al., 2021*; *Motta, Stecula & Farhart, 2020*), there are some others that frame it otherwise (*Pedersen & Favero, 2020*; *Piltch-Loeb et al., 2021*). Thus, they cancel each other out and as a result, put the overall tones in a relatively neutral position. In other words, if we put all media in one basket, their news coverage portrays the pandemic with both positive and negative nuances, resulting in a balanced and rather neutral overall tone. The only exception, however, is for death-related news, in which East media used a significantly more negative tone. Indeed, despite death being unambiguously considered by most as negative news and the fact that negative news attracts more readers (*Trussler & Soroka, 2014*), it is still intriguing to investigate why only this particular topic of COVID-19 was framed with a more negative tone by the mass media in the East.

Regardless of the tone used by the media in their coverage of COVID-19, it is still largely up to the public to express either positive or negative reactions. Even though the public in both regions tend to have negative reactions in general, our findings show that the public in the West has significantly more negative reactions than the public in the East, as shown in Table 3. In other words, people in the West region appear to be perceiving the COVID-19 pandemic as more devastating than those in the East region. One possible explanation for this phenomenon is related to the cumulative cases and deaths per million people, as illustrated in Figs. 2 and 3. In this respect, the top five countries with the highest number of cumulative cases and deaths in our dataset were all located in the West region.

Another possible explanation is individualist culture, which appears to influence how the public in the Western region reacted to COVID-19 security measures and policies (*Chen, Frey & Presidente, 2021*; *Leonhardt & Pezzuti, 2022*). The pandemic has urged people, regardless of their health status, to take various preventive measures to stop the

virus' spread and mitigate casualties. Perhaps these policies have been perceived as threats to freedom and liberty by a sizable portion of the West's population, who predominantly hold individualistic values and whose compliance is primarily motivated by self-interest and a sense of individual obligation to support authorities (*Banker & Park, 2020*; *Favero & Pedersen, 2020*; *Murphy et al., 2020*). As such, they might find preventive measures such as social distancing, mask-wearing, or vaccinations as annoying, inconvenient, or even a threat to their personal freedom (*Johnston, 2022*; *Rains et al., 2022*; *Young et al., 2022*).

Meanwhile, for many people in the East, who predominantly have collectivistic values (*Hofstede, 2011*; *Hofstede, Hofstede & Minkov, 2005*), the said policies might have been perceived as ways to care for each other in their community. They do not mind sacrificing their individual convenience for the sake of their community, so that there is no difference between their self-interest and prosocial motivations behind their compliance (*Chua et al., 2021*). It can also explain why the public in the East showed a more positive reaction to COVID-19 vaccine-related news compared to the public in the West (see Fig. 6). Vaccines not only protect oneself but also others who are even more vulnerable. Indeed, this more positive reaction does not necessarily translate to a higher vaccination rate, as shown in Fig. 4. However, it is arguably due to socio-economic factors that most Eastern countries in this study do not have the same access to COVID-19 vaccines as their Western counterparts (*Acharya, Ghimire & Subramanya, 2021*). Evidently, when they have the means, countries in the East, such as Singapore and China, were among the top three countries in our dataset with the highest vaccination rates. Indeed, such political factors cannot be excluded from the equation. It is also possible that the public in the East region was afraid of the negative consequences imposed by their governments *e.g.*, (*Dyer, 2021*) should they decide to not get vaccinated. It could also prevent or even stop them from making negative comments on social media about COVID-19 news.

In addition, considering that Facebook users with certain demographic backgrounds (*i.e.*, female, older, less educated) tend to give more 'likes' than others (*Hong, Chen & Li, 2017*), it could be the case that users with certain demographics were also more likely to give certain special reactions to the same news article compared to users of other demographics. Perhaps users belong to the former group are more familiar with the Facebook features, have more time to spend on the platform, or possess more means to access the site. Meanwhile, users belong to the latter group tend not to give any special reactions, probably, because they are less knowledgeable of doing it, have less time to spend on Facebook, or deliberately choose not to leave any digital trace on the platform. Since we do not have such demographics data and are restricted to acquire ones by Facebook and CrowdTangle, we could not confirm if our proposed explanation holds in this context.

Study limitations

With the strengths of using a larger scale of data and employing an innovative method of computational social science, there are some limitations of this study that are worth mentioning. We thus suggest readers be aware of the following limitations when interpreting such results.

First, we only focus on news articles written in English on Facebook with no other topics as a sentiment baseline. We could not include other languages since they have not been supported by the current package used for the main analysis *i.e.*, "vaderSentiment" (*Hutto & Gilbert, 2014*). We were less able to include another topic as the baseline due to our current resources from CrowdTangle prevents us from doing so. This limitation then narrows the countries included in this study and makes Vader, to some extent, as the baseline tool. For the West region, all of them are English speaking countries. For the East region, only a few of them use English as one of their official languages. This limitation also explains why the media in the West region had more Facebook posts and received more special reactions than media in the East. In addition, it is not a given that the news delivered in English in these non-English-speaking Eastern countries shares the same tone or receives the same public reactions as most of the news delivered in their native languages.

Second, we do not have users' demographic information that can help explain the disparity in the public reactions between West and East regions in much greater detail. For one, the information was not provided by CrowdTangle. Indeed, we could do data scraping and additional procedures on our own to get the demographic data. However, it would be against Meta's and Facebook's policies, which forbid data mining in any form, including for academic purposes (*Clark, 2021*; *Vittorio, 2021*).

Third, we only examined associations between variables of interest. Thus, such significant findings or larger effect sizes should not be interpreted as causality. This limitation also applies to the explanation regarding the disparity between public reactions in the West and East. Even though our arguments are supported by past research, we have yet to validate them empirically, for instance, by using statistical procedures.

Finally, the facts that Facebook penetrations were not the same across the countries analyzed in this study (*Dixon, 2022*) and different mass media might use Facebook differently give a further precaution for readers when making such generalizations. In this regard, the results, albeit derived from a relatively huge dataset, might not necessarily reflect the whole population in the country under study. Likewise, the results may not always be representative of the entire media system in the regions under study. Some mass media might use Facebook more frequently and thus have a greater chance to be in our dataset than other media. Unfortunately, this current study was less able to give detailed pictures of the types of media that were specifically examined, except the fact that they were all digital mass media sharing news articles both in their own website and Facebook.

Directions for future research

Software-wise, we encourage future developers to address the current tool's limitations by, for instance, developing sentiment analysis packages in other languages that are compatible with Vader to allow for meaningful comparison with its English counterpart. In addition, we recommend future studies to use other points of views in examining the differences or similarities between news tones and public reactions in the West and East regions. For instance, it would be interesting to investigate the comparisons through political leanings of the media or the readers. It would also be more fruitful to learn

demographic and examine other psychological factors including individualism/collectivism personality traits that may influence such reactions. Finally, we also motivate future researchers to do a replication study, in particular with smaller units of analysis (*e.g.*, country level) or by using a sentiment baseline. Doing so, we believe, will further validate the study findings.

CONCLUSION

This study aims to investigate COVID-19 news sentiment by looking into the differences in tone used by the mass media in the West and East in their COVID-19 pandemic news coverage in English, as well as public reactions to it, and if there is any correlation between the two. Through analyzing archival data of news posted by mass media on Facebook with the help of a sentiment analysis software, Vader, we discovered that, in aggregate, news in both West and East media generally contained a relatively similar level of rather neutral tone. However, people in Western countries reacted significantly more negatively to COVID-19 news than their Eastern counterparts. Interestingly, we found none to strong correlations between the news tones and public reactions in both regions suggesting case-by-case relationships depending on the topics. We then argue that this disparity might be driven by the individualism *vs.* collectivism cultural differences along with different socio-economic-political contexts between countries in the West and East regions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ahmad R. Pratama conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Firman M. Firmansyah conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The summarized and anonymized data along with the code used to processed it are available at GitHub: https://github.com/ahmadrafie/fbcovid19newssentiment.

Per CrowdTangle's policy, the raw dataset cannot be shared publicly.

The dataset can be retrieved by a CrowdTangle user with the following search parameters: Keyword = covid-19; Platform = Facebook; Account Type = Pages; Timeframe

= March 1, 2020 - February 28, 2022; Post Type = Links; Language = English, List = Asia Media, Australia Media, Canada Media, UK Media, and US General Media (all managed and maintained by the CrowdTangle team) with all other default settings.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1111#supplemental-information.

REFERENCES

Acharya KP, Ghimire TR, Subramanya SH. 2021. Access to and equitable distribution of COVID-19 vaccine in low-income countries. *NPJ Vaccines* **6(1)**:54 DOI 10.1038/s41541-021-00323-6.

Ahmad Kamboh S, Ittefaq M, Sahi AA. 2022. Journalistic routines as factors promoting COVID-19 vaccine hesitancy in Pakistan. *Third World Quarterly* **43(1)**:278–287 DOI 10.1080/01436597.2021.1995713.

Assistant Secretary for Health. 2020. *Community-based testing sites for COVID-19*. HHS.gov; US Department of Health and Human Services. *Available at https://www.hhs.gov/coronavirus/community-based-testing-sites/index.html*.

Banker S, Park J. 2020. Evaluating prosocial COVID-19 messaging frames: evidence from a field study on Facebook. *SSRN Electronic Journal* **148(3)**:1005 DOI 10.2139/ssrn.3684901.

Bryant C. 2022. Researchpy Documentation. *Available at https://researchpy.readthedocs.io/en/latest/*.

Burton L, Delvecchio E, Germani A, Mazzeschi C. 2021. Individualism/collectivism and personality in Italian and American Groups. *Current Psychology* **40(1)**:29–34 DOI 10.1007/s12144-019-00584-4.

Chaudhary FA, Ahmad B, Khalid MD, Fazal A, Javaid MM, Butt DQ. 2021. Factors influencing COVID-19 vaccine hesitancy and acceptance among the Pakistani population. *Human Vaccines & Immunotherapeutics* **17(10)**:3365–3370 DOI 10.1080/21645515.2021.1944743.

Chen Y, Biswas MI. 2022. Impact of national culture on the severity of the COVID-19 pandemic. Epub ahead of print 24 February 2022. *Current Psychology* 1–14 DOI 10.1007/s12144-022-02906-5.

Chen C, Frey CB, Presidente G. 2021. Culture and contagion: individualism and compliance with COVID-19 policy. *Journal of Economic Behavior & Organization* **190(5)**:191–200 DOI 10.1016/j.jebo.2021.07.026.

Chua CE, Kew GS, Demutska A, Quek S, Loo EXL, Gui H, Wong S, Lau HX, Low EXS, Loh TL, Ooi SL, Hung ECW, Rahman MM, Ghoshal UC, Wong H, Cheung CKY, Syam AF, Tan N, Xiao Y, Liu J-S, Lu F, Chen C-L, Lee YY, Maralit RM, Kim Y-S, Oshima T, Miwa H, Pang J, Siah KTH. 2021. Factors associated with high compliance behaviour against COVID-19 in the early phase of pandemic: a cross-sectional study in 12 Asian countries. *BMJ Open* **11(8)**:e046310 DOI 10.1136/bmjopen-2020-046310.

Clark M. 2021. Research cannot be the justification for compromising people's privacy. Meta. *Available at https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/*.

CrowdTangle Team. 2019. *CrowdTangle*. CrowdTangle. *Available at https://crowdtangle.com/*.

Delacre M, Lakens D, Leys C. 2017. Why psychologists should by default use welch's *t*-test instead of student's *t*-test. *International Review of Social Psychology* **30(1)**:92 DOI 10.5334/irsp.82.

Dixon S. 2022. *Facebook users by country 2021*. Statista. *Available at https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/*.

Dyer O. 2021. Covid-19: singaporeans who remain "unvaccinated by choice" must fund own treatment, says government. *BMJ* **375**:n2748 DOI 10.1136/bmj.n2748.

Favero N, Pedersen MJ. 2020. How to encourage "Togetherness by Keeping Apart" amid COVID-19? The ineffectiveness of prosocial and empathy appeals. *Journal of Behavioral Public Administration* **3(2)**:1–18 DOI 10.30636/jbpa.32.167.

Firmansyah FM, Jones JJ. 2019. Did the black panther movie make blacks blacker? Examining black racial identity on twitter before and after the black panther movie release. *Social Informatics* **11864**:66–78 DOI 10.1007/978-3-030-34971-4.

Firmansyah FM, Pratama AR. 2021. Anonymity in COVID-19 online donations: a cross-cultural analysis on fundraising platforms. *Advances in Information and Communication* **2**:34–47 DOI 10.1007/978-3-030-73103-8_3.

Freeman C, Alhoori H, Shahzad M. 2020. Measuring the diversity of facebook reactions to research. *Proceedings of the ACM on Human-Computer Interaction* **4(GROUP)**:1–17 DOI 10.1145/3375192.

Fritz CO, Morris PE, Richler JJ. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General* **141(1)**:2–18 DOI 10.1037/a0024338.

Hofstede G. 2011. Dimensionalizing cultures: the hofstede model in context. *Online Readings in Psychology and Culture* **2(1)**:8 DOI 10.9707/2307-0919.1014.

Hofstede G, Hofstede GJ, Minkov M. 2005. *Cultures and organizations: software of the mind.* Vol. 2. New York: McGraw-Hill.

Hofstede G, McCrae RR. 2004. Personality and culture revisited: linking traits and dimensions of culture. *Cross-Cultural Research: Official Journal of the Society for Cross-Cultural Research/ Sponsored by the Human Relations Area Files, Inc* **38(1)**:52–88 DOI 10.1177/1069397103259443.

Hong C, Chen Z, Li C. 2017. "Liking" and being "liked": how are personality traits and demographics associated with giving and receiving likes on Facebook? *Computers in Human Behavior* **68**:292–299 DOI 10.1016/j.chb.2016.11.048.

Hossain A, Wahab JA, Khan MSR. 2022. A computer-based text analysis of Al Jazeera, BBC, and CNN news shares on Facebook: framing analysis on Covid-19 issues. *SAGE Open* **12(1)**:21582440211068497 DOI 10.1177/21582440211068497.

Hutto C, Gilbert E. 2014. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 216–225.

Johnston RD. 2022. Vaccine resistance and history in the land of liberty [Review of *Vaccine Resistance and History in the Land of Liberty*]. *American Journal of Public Health* **112(2)**:227–228 DOI 10.2105/AJPH.2021.306583.

Leonhardt JM, Pezzuti T. 2022. Vaccination acceptance across cultures: the roles of collectivism, empathy, and homophily. *Journal of International Marketing* **30(2)**:13–27 DOI 10.1177/1069031X211073179.

Maaravi Y, Levy A, Gur T, Confino D, Segal S. 2021. "The tragedy of the commons": how individualism and collectivism affected the spread of the COVID-19 pandemic. *Frontiers in Public Health* **9**:627559 DOI 10.3389/fpubh.2021.627559.

Mach KJ, Salas Reyes R, Pentz B, Taylor J, Costa CA, Cruz SG, Thomas KE, Arnott JC, Donald R, Jagannathan K, Kirchhoff CJ, Rosella LC, Klenk N. 2021. News media coverage of COVID-19 public health and policy information. *Humanities and Social Sciences Communications* **8(1)**:1–11 DOI 10.1057/s41599-021-00900-z.

Massola J. 2020. More than a month's salary for a COVID test? Welcome to Indonesia. *The Sydney Morning Herald. Available at https://www.smh.com.au/world/asia/more-than-a-months-salary-for-a-covid-test-welcome-to-indonesia-20200714-p55bvq.html*.

Medina LM, Rodriguez JR, Sarmiento PJD. 2021. Shaping public opinion through the lens of agenda setting in rolling out COVID-19 vaccination program. *Journal of Public Health* **43(2)**: e389–e390 DOI 10.1093/pubmed/fdab100.

Motta M, Stecula D, Farhart C. 2020. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the U.S. *Canadian Journal of Political Science/Revue Canadienne de Science Politique* **53(2)**:335–342 DOI 10.1017/S0008423920000396.

Murphy K, Williamson H, Sargeant E, McCarthy M. 2020. Why people comply with COVID-19 social distancing restrictions: Self-interest or duty? *The Australian & New Zealand Journal of Criminology* **53(4)**:477–496 DOI 10.1177/0004865820954484.

Oesper L, Merico D, Isserlin R, Bader GD. 2011. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code for Biology and Medicine* **6(1)**:7 DOI 10.1186/1751-0473-6-7.

Ogbodo JN, Onwe EC, Chukwu J, Nwasum CJ, Nwakpu ES, Nwankwo SU, Nwamini S, Elem S, Iroabuchi Ogbaeja N. 2020. Communicating health crisis: a content analysis of global media framing of COVID-19. *Health Promotion Perspectives* **10(3)**:257–269 DOI 10.34172/hpp.2020.40.

Oliveira DJS, Bermejo PHDS, dos Santos PA. 2017. Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics* **14(1)**:34–45 DOI 10.1080/19331681.2016.1214094.

Oueslati O, Khalil AIS, Ounelli H. 2018. Sentiment analysis for helpful reviews prediction. *International Journal of Advanced Trends in Computer Science and Engineering* **7(3)**:34–40 DOI 10.30534/ijatcse/2018/02732018.

Pedersen MJ, Favero N. 2020. Social distancing during the COVID-19 pandemic: who are the present and future non-compliers? *Public Administration Review* **80(5)**:805–814 DOI 10.1111/puar.13240.

Piltch-Loeb R, Savoia E, Goldberg B, Hughes B, Verhey T, Kayyem J, Miller-Idriss C, Testa M. 2021. Examining the effect of information channel on COVID-19 vaccine acceptance. *PLOS ONE* **16(5)**:e0251095 DOI 10.1371/journal.pone.0251095.

Pratama AR. 2022. Sentiment analysis of Facebook posts through special reactions: the case of learning from home in Indonesia Amid COVID-19. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika* **8(1)**:83–92 DOI 10.26555/jiteki.v8i1.23615.

Rains SA, Harber P, Warner EL, Leroy G. 2022. Public responses to COVID-19 mask mandates: examining pro and anti-Mask anger in tweets before and after state-level mandates. *Communication Monographs* **12(1)**:1–19 DOI 10.1080/03637751.2022.2043555.

Rajkumar RP. 2021. The relationship between measures of individualism and collectivism and the impact of COVID-19 across nations. *Public Health in Practice (Oxford, England)* **2**:100143 DOI 10.1016/j.puhip.2021.100143.

Rianto A, Pratama AR. 2021. Sentiment analysis of COVID-19 vaccination posts on Facebook in Indonesia with CrowdTangle. *Jurnal Riset Informatika* **3(4)**:353–362 DOI 10.34288/jri.v3i4.282.

Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, Hasell J, Macdonald B, Beltekian D, Roser M. 2020. *Coronavirus Pandemic (COVID-19).* OurWorldInData.org. *Available at https://ourworldindata.org/coronavirus*.

Rozado D. 2021. Prevalence in news media of two competing hypotheses about COVID-19 origins. *Social Sciences* **10(9)**:320 DOI 10.3390/socsci10090320.

Statista. 2022. *Most popular social networks worldwide as of January 2022, ranked by number of monthly active users*. Statista. *Available at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users*.

Tian Y, Galery T, Dulcinati G, Molimpakis E, Sun C. 2017. Facebook sentiment: reactions and emojis. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 11–16.

Triandis HC. 2001. Individualism-collectivism and personality. *Journal of Personality* **69(6)**:907–924 DOI 10.1111/1467-6494.696169.

Trussler M, Soroka S. 2014. Consumer demand for cynical and negative news frames. *The International Journal of Press/Politics* **19(3)**:360–379 DOI 10.1177/1940161214524832.

Vittorio A. 2021. *Facebook move against NYU research stirs scraping policy feud*. Bloomberg Law. *Available at https://news.bloomberglaw.com/privacy-and-data-security/facebook-move-against-nyu-research-stirs-scraping-policy-feud*.

Wu K. 2021. Agenda-setting in cross-national coverage of COVID-19: an analysis of elite newspapers in US and China with topic modeling. *Online Journal of Communication and Media Technologies* **11(4)**:e202116 DOI 10.30935/ojcmt/11083.

Young DG, Rasheed H, Bleakley A, Langbaum JB. 2022. The politics of mask-wearing: political preferences, reactance, and conflict aversion during COVID. *Social Science & Medicine* **298(3)**:114836 DOI 10.1016/j.socscimed.2022.114836.

Zhao E, Wu Q, Crimmins EM, Ailshire JA. 2020. Media trust and infection mitigating behaviours during the COVID-19 pandemic in the USA. *BMJ Global Health* **5(10)**:e003323 DOI 10.1136/bmjgh-2020-003323.

PeerJ Computer Science

Resilience of political leaders and healthcare organizations during COVID-19

Manmeet Kaur Baxi[1], Joshua Philip[2] and Vijay Mago[1]

[1] Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
[2] Superior Collegiate and Vocational Institute, Thunder Bay, Ontario, Canada

ABSTRACT

This study assesses the online societal association of leaders and healthcare organizations from the top-10 COVID-19 resilient nations through public engagement, sentiment strength, and inclusivity and diversity strength. After analyzing 173,071 Tweets authored by the leaders and health organizations, our findings indicate that United Arab Emirate's Prime Minister had the highest online societal association (normalized online societal association: 1.000) followed by the leaders of Canada and Turkey (normalized online societal association: 0.068 and 0.033, respectively); and among the healthcare organizations, the Public Health Agency of Canada was the most impactful (normalized online societal association: 1.000) followed by the healthcare agencies of Turkey and Spain (normalized online societal association: 0.632 and 0.094 respectively). In comparison to healthcare organizations, the leaders displayed a strong awareness of individual factors and generalized their Tweets to a broader audience. The findings also suggest that users prefer accessing social media platforms for information during health emergencies and that leaders and healthcare institutions should realize the potential to use them effectively.

Subjects Natural Language and Speech, Social Computing, Text Mining, Sentiment Analysis
Keywords Social media, Twitter, COVID-19, User engagement, Content analysis, Sentiment strength, Inclusivity and diversity strength, Crisis communication

Submitted 3 May 2022
Accepted 7 September 2022
Published 7 October 2022

Corresponding author
Manmeet Kaur Baxi,
mbaxi@lakeheadu.ca

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj-cs.1121**

Copyright
2022 Baxi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

The exponential spread of the 2019 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has flooded various social media platforms (SMPs) with a plethora of information about the disease, pandemic trajectory, influence on human fatalities, and global and regional consequences for the governments and health organizations (*Gates, 2020*). SMPs such as Twitter, Facebook and Instagram have become the norm for broadcasting and acquiring pandemic-related information by leaders and healthcare organizations. Researchers have demonstrated an interest in examining and interpreting the social media data of leaders and healthcare organizations across various SMPs by evaluating their Twitter usage through content analysis and the change in their number of followers (*Haman, 2020*; *Rufai & Bunce, 2020*).

Political leaders are followed on Twitter for a number of reasons, including convenience, expressiveness, knowledge and sociability (*Parmelee & Bichard, 2011*). According to research, 70% of healthcare institutions in the United States also utilize

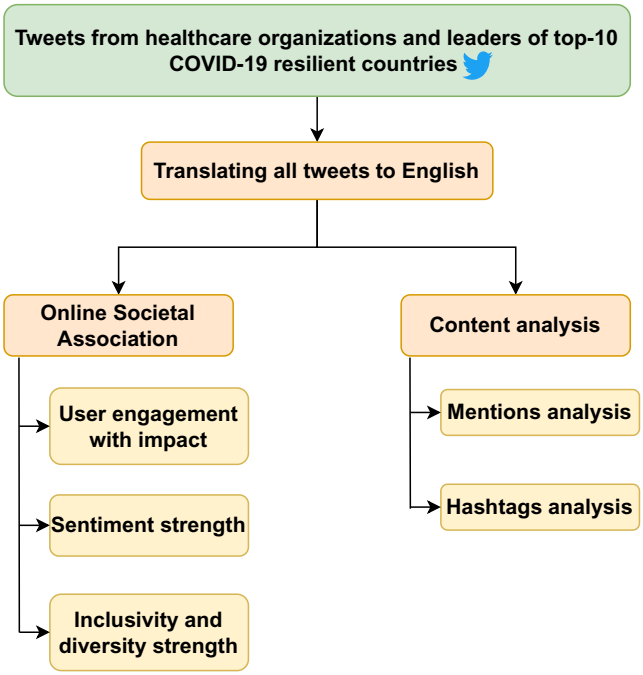


Figure 1 **Overall research framework.** Full-size DOI: 10.7717/peerj-cs.1121/fig-1

SMPs, and their social media presence influences 57% of clients' decisions about where to seek medical care (*Peck, 2014*). Thus, it is necessary to investigate the online societal association of leaders and healthcare organizations on the citizens during a crisis situation. The importance of studying the qualitative factors to determine the online societal association of social media on Government to Citizen interactions has been highlighted previously (*Bonsón et al., 2012*; *Norris & Reddick, 2013*). Therefore, the objective of this study is to quantify the online societal association of leaders and healthcare organizations from the top-10 COVID-19 resilient nations by analyzing the following factors–user engagement, sentiment strength, and the inclusivity and diversity of various communities in the Tweets authored by them. Figure 1 depicts the overall research framework.

Understanding which information (content, type) appeals to the audience the most might be an effective way of amplifying the involvement (*Bonsón, Royo & Ratkai, 2015*; *Bonsón & Bednárová, 2018*), and hence can help in perpetuating a regular conversation with the public, addressing their concerns, recommendations, and desires, and thus assist in pacifying them, building trust, and fighting through crisis situations together. On this rationale, this article calculates the influence of leaders and healthcare organizations, and the engagement they receive during COVID-19. Furthermore, people, especially political leaders and healthcare organizations, communicate their opinions and attitudes–which are generally termed as '*sentiment*' through several SMPs (*Dang-Xuan et al., 2013*). However, it is ambiguous how sentiment and the references to different communities might influence information dynamics in a social-media setting. Therefore, it is critical to evaluate the collective influence of sentiment, inclusion and diversity along with the public engagement on the leaders' and healthcare organizations' online societal association.

This study offers a thorough assessment of the leaders and healthcare organizations, including both qualitative and quantitative insights regarding their online behaviour. This type of hybrid study has yielded promising findings in the past as well (*Sampieri, 2018*). Our research examines the various metrics of their Tweets (likes, replies, retweets, and quotes) as well as utilizes statistical methodologies to compute sentiment strength, inclusiveness, and diversity strength. It is qualitative in nature as we analyze the Tweet content to see whether there are any parallels to real-life events and we try to quantify the levels of engagement and online societal association. As a result, the purpose of this study is to explore how audience engagement, sentiment, inclusion and diversity strength may assist leaders and healthcare organizations develop a trustworthy relationship with the public, gathering support for policies that limit the spread of COVID-19, and overcoming the crisis situations. The key findings of the research are:

1. Amongst politicians, United Arab Emirates' Prime Minister had the highest online societal association,
2. The Canadian public health agency demonstrated a prominent level of online societal association amid the healthcare organizations, and,
3. Individual aspects of the online societal association were better understood by leaders than by healthcare organizations, because the latter had lower levels of audience engagement, targeted limited groups, and had relatively low sentiment strength, as seen by the results.

The remainder of the article is structured as follows: Related work section discusses the summary of the past research relevant to the usage of SMPs by leaders and healthcare organizations. Dataset and the statistical methods used to compute the online societal association are discussed in the Methodology section. The results and implications of this study are discussed in Results, followed by the Discussion & Conclusion section, where the key inferences and further research directions are discussed.

RELATED WORK

Leaders and healthcare organizations utilize social media platforms (SMPs) including Facebook, Twitter, Instagram, and Reddit for election campaigns, broadcasting public health information, announcing significant events, and improving public relations (*Bhattacharya, Srinivasan & Polgreen, 2017*; *Bertot et al., 2010*; *Bonsón et al., 2012*; *Chun et al., 2010*).

In lieu of the use of SMPs for the candidate and audience engagement on Twitter, and election campaigns for the 2013 and 2016, Australian federal elections was compared (*Bruns & Moon, 2018*), while the use of Facebook for strategic campaigning in the 2008 and 2012 US Presidential elections was analyzed (*Borah, 2016*). Additionally, the use of Instagram as an advertising tool for Swedish elections and YouTube for the electoral campaign of European Parliamentary elections was examined (*Russmann & Svensson, 2017*; *Vesnic-Alujevic & Van Bauwel, 2014*). SMPs, according to earlier studies (*Bertot et al., 2010*; *Bonsón et al., 2012*; *Chun et al., 2010*) may assist in improving governance

transparency, engagement, and accountability. Furthermore, using a variety of examples, researchers have highlighted the impact on several aspects of public interaction through SMPs (*Gruzd & Roy, 2016*; *Hollebeek, Glynn & Brodie, 2014*; *Ríos, Benito & Bastida, 2017*). These platforms have been cited by several academicians (*Bonsón, Perea & Bednárová, 2019*; *Bonsón & Ratkai, 2013*; *Bonsón, Royo & Ratkai, 2017*; *Gruzd, Lannigan & Quigley, 2018*; *Sahly, Shao & Kwon, 2019*; *Siebers, Gradus & Grotens, 2019*) as an important tool for expanding the social reach and better understanding the audience. However, previous research has also demonstrated that the sentiment, emotion, and diversity of Tweets, can mitigate public engagement and persuade the audience (*Bhat et al., 2021*; *Sandoval-Almazan & Valle-Cruz, 2018*; *Jünger & Fähnrich, 2020*; *Qudar, Bhatia & Mago, 2021*; *Singhal, Baxi & Mago, 2022*). Therefore, it is clear that in past few years, political leaders from a variety of nations have actively employed SMPs for both their election campaigns and to comprehend their audience.

As seen recently, SMPs have also been used by healthcare professionals and organizations as a communication tool to promote healthy habits, share announcements, disseminate awareness, motivate the patients on the way to their recovery, support emergency response, and eventually boost readiness during exceptional circumstances (*Benetoli, Chen & Aslani, 2018*; *Househ, 2013*; *Li & Sakamoto, 2014*; *Ventola, 2014*; *Merchant & Lurie, 2020*; *Shah et al., 2020*; *Patel et al., 2019*, *2020*). The idea of SMPs having a positive influence on public awareness and behavioural changes by disseminating succinct information to specified audiences was explored (*Al-Dmour et al., 2020*). Further, to evaluate public reactions during the epidemic, topic identification and sentiment analysis was utilized to examine the change of public attitude over time in relation to the published news, and reddit posts (*Garcia & Berton, 2021*; *Melton et al., 2021*). Also, a previous study has identified factors associated with the levels and duration of engagement, for the Facebook accounts of U.S. Federal health agencies (*Bhattacharya, Srinivasan & Polgreen, 2017*). Furthermore, researchers have investigated the influence of world leaders during the COVID-19 pandemic and how Twitter was used to swiftly transmit information to the public (*Haman, 2020*; *Rufai & Bunce, 2020*). However, there has not been an extensive analysis to understand and examine the online societal association of leaders and healthcare organizations during COVID-19, considering different factors like user engagement, sentiment strength, and, inclusivity and diversity strength, during a health emergency, despite the physical distancing and lockdown measures, which is hence the focus of this work.

METHODOLOGY

Dataset

Twitter Academic Research API v2 (https://developer.twitter.com/en/products/twitter-api/academic-research) was utilized to retrieve the information of the political leaders' and health organizations' Tweets. A total of 173,071 Tweets were collected and analyzed from December 1, 2019, to December 31, 2021. The dataset was curated based on the Bloomberg COVID-19 Resilience Ranking (https://www.bloomberg.com/graphics/covid-resilience-ranking/), as of January 8, 2022, at 5 p.m. EST, selecting the health organizations and

Table 1 Distribution of Tweets for the selected user accounts.

Account type	Name (Twitter handle)	Country	Number of Tweets
Leader (President or Prime Minister)	Sebastián Piera (@sebastianpinera) (President)	Chile	622
	Micheál Martin (@PresidentIRL, @MichealMartinTD)	Ireland	3,641
	Mohammed bin Rashid Al Maktoum (@HHShkMohd)	U.A.E	839
	Sanna Marin (@MarinSanna)	Finland	2,007
	Justin Trudeau (@JustinTrudeau, @CanadianPM)	Canada	13,778
	Iván Duque (@IvanDuque) (President)	Colombia	7,059
	Recep Tayyip Erdoğan (@RTErdogan) (President)	Turkey	1,943
	Pedro Sánchez (@sanchezcastejon)	Spain	4,290
	Magdalena Andersson (@SwedishPM)	Sweden	282
	Boris Johnson (@BorisJohnson)	United Kingdom	2,335
	Total		36,796
Health Organization/Health Minister	Ministerio de Salud (@ministeriosalud)	Chile	39,401
	HSELive (@HSELive), Department of Health (@roinnslainte)	Ireland	18,332
	Ministry of Health and Prevention, U.A.E. (@mohapuae)	U.A.E	8,424
	Ministry of Social Affairs and Health (@MSAH_News)	Finland	1,009
	Health Canada and PHAC (@GovCanHealth)	Canada	38,715
	Ministry of Health and Social Protection of Colombia (@MinSaludCol)	Colombia	11,346
	Ministry of Health of the Republic of Turkey (@saglikbakanligi)	Turkey	4,119
	Ministry of Health (@sanidadgob)	Spain	8,595
	The Public Health Agency of Sweden (@Folkhalsomynd)	Sweden	701
	UK Health Security Agency (@UKHSA)	United Kingdom	5,633
	Total		136,725

leaders of the top-10 COVID-19 resilient countries. The COVID-19 Resilience Ranking is a monthly impression of the countries handling the virus most effectively, with the least social and economic disruption. The ranking is calculated based on the factors of virus containment, quality of healthcare, vaccination coverage, overall mortality and progress towards restarting travel. The timeline was chosen to include the outbreak COVID-19 to the vaccination period of the pandemic. Official health organizations of the respective countries and personal accounts of the political leaders were analyzed in this specific study. This provides an opportunity to get a sense of the contrasting dynamics between the accounts; to truly encapsulate the online societal association on the particular country.

The collected Tweets spanned across 19 different languages and were translated to English using the Neural Machine Translation (NMT) models from the Tatoeba Translation Challenge, which consists of NMT models trained on a compressed dataset of over 500 GB, encompassing 2,961 language pairings, and 555 languages (*Tiedemann, 2020*). For this study, each Twitter account is referred to as a *user* and the type of account (*i.e.*, leaders, health organizations) is referred to as a *user group*. The details of the Tweets authored by each of the selected users in the order of their COVID-19 Resilience ranking

(*i.e.*, from the best country to live in during COVID-19, like Chile, to the good ones, like United Kingdom) can be found in Table 1.

Online societal association

The online societal association (denoted by, *onlineSocietalAssociation*) is defined as the product of engagement per day with user impact ($dailyAvgEng_{user}$), sentiment strength (*sentiStrength*), and inclusivity and diversity strength (*iDStrength*) in the user's Tweets as in Eq. (1). Each criterion is given the same weight since they have all been scaled to account for bias, and all of these parameters come together to generate a Tweet which is addressed to the general audience amongst the network of healthcare organizations and leaders of the top-10 COVID-19 resilient countries in the Twitter ecosystem. Further details of the variables can be found in the following subsections.

$$onlineSocietalAssociation = dailyAvgEng_{user} * sentiStrength * iDStrength \tag{1}$$

Engagement with impact

The engagement per day is the measure of the social interaction of the post, including the likes, replies, retweets and quotes. The engagement per day represents the relationship between the followers and the user, and the resonation of their Tweets. Twitter defines engagement rate, as the ratio of engagements to impressions: $\frac{Engagement}{Impressions} \times 100$. The engagements are defined as an aggregate of interactions of a Tweet–retweets, replies, follows, likes, links, cards, hashtags, embedded media, profile photo, username or Tweet expansion. The impressions account for times a user has observed a particular Tweet in their search results or timeline (Twitter Account Activity Analytics–Engagement, Impressions) (https://help.twitter.com/en/managing-your-account/using-the-Tweet-activity-dashboard). This study analyzes only public metrics such as the count of likes, replies, retweets and quotes-as a result of the limitations of Twitter API.

To evaluate a user's engagement ($dailyAvgEng_{user}$, Eq. (2)); firstly, their Tweet-wise engagement ($dailyAvgEng_{(Tweet,user)}$, Eq. (3)) is calculated by multiplying the user impact ($impact_{user}$) and average engagement per day for a Tweet, ($dailyAvgEng_{Tweet}$), followed by taking an average of the Tweet-wise engagement ($dailyAvgEng_{(Tweet,user)}$) for the user.

$$dailyAvgEng_{user} = \frac{\sum dailyAvgEng_{(Tweet,user)}}{totalTweets_{user}} \tag{2}$$

$$dailyAvgEng_{(Tweet,user)} = dailyAvgEng_{Tweet} \times impact_{user} \tag{3}$$

The impact of a user ($impact_{user}$, Eq. (4)) is quantified based upon the hyperbolic tangent function (tanh) of followers, the total number of Tweets, following, public lists and profile age. The *listedCount* is the total amount of public lists of a user. Lists indicate popularity–generally revolving around the concept that other users are engaged with one's content. Furthermore, $log_{10}\left(\frac{\sqrt{followers}}{following}\right)$ is the followers to following ratio, indicating the general nature of the account. The ratio is within a base-10 log to elude outlier values. The *totalTweetCount* is the number of Tweets from the account during our data collection

timeframe. The *profileAge* represents the number of days between the profile creation date to December 31, 2021; the last analysis day. Because a freshly joined user with more followers would be more influential than a previously joined user with fewer followers, the square of a user's profile age has been deemed inversely proportional to the user's impact.

$$impact_{user} = \frac{\tanh\left(\log_{10}\left(\frac{\sqrt{followers}}{following}\right) \times listedCount \times TweetCount\right)}{(profileAge)^2} \tag{4}$$

To quantify the average engagement per day ($dailyAvgEng_{Tweet}$, Eq. (5)), the collated number of likes, replies, retweets, quotes and Tweets per day, from December 1, 2019, to December 31, 2021, is used. Furthermore, the *dailyTweetCount* are multiplied by 4 (equal to the number of variables in the numerator).

$$dailyAvgEng_{Tweet} = \frac{likes + replies + retweets + quotes}{4 * dailyTweetCount} \tag{5}$$

To standardize the shifting values of average engagement, we calculate the Exponential Moving Average with a 151-day window span[1], eliminate outliers using $z - score$ and smoothen the average engagement per day to the 8th degree using the Savitzky Golay filter[2].

[1] A grid-search analysis was performed to find the best value.

[2] A grid-search analysis was performed to find the best value.

Sentiment strength

To quantify the strength of sentiment for every user, we first calculate the sentiments of all the Tweets collected for our analysis using CardiffNLP's '*twitter-roberta-base-sentiment*' model, which is trained on a 60 million Twitter *corpus*, and then calculate the sentiment strength for every user as mentioned in Eq. (6), *i.e.*, based on the sentiment category with the maximum number of Tweets for that day, followed by assigning the sentiment score based on the sentiment: 10^{-6} for neutral, the ratio of the count of positive Tweets to total Tweets for positive, and the negation of the ratio of the count of negative Tweets to the total Tweets for negative sentiment.

$$sentiStrength = \begin{cases} 10^{-6} & ; maxSentimentScore(Tweets) = neutral \\ \frac{count(positiveTweets)}{totalTweets} & ; maxSentimentScore(Tweets) = positive \\ -\frac{count(negativeTweets)}{totalTweets} & ; maxSentimentScore(Tweets) = negative \end{cases} \tag{6}$$

Inclusivity and diversity strength

We assessed the inclusivity and diversity in the Tweets of the users (denoted by, *iDStrength*) by computing the usage of different keywords pertaining to various communities from the countries selected for our analysis. The keywords were selected based on gender, age, cultural inferences, ethnicity, and employment sectors of each of these countries. The detailed list of keywords can be found in the GitHub repository

(https://github.com/manmeetkaurbaxi/Societal-Impact-on-Twitter). The usage frequency for each of these keywords is calculated for all users with respect to the total number of Tweets from that user, as given in Eq. (7). If there exists a user who has not referred to any community in their Tweets, a default value of 10^{-6} is assigned.

$$iDStrength = \begin{cases} 10^{-6} & ;count(communityMentionTweets) = 0 \\ \frac{count(communityMentionTweets)}{totalTweets}; & otherwise \end{cases} \quad (7)$$

Content analysis

The Tweets of all the users were analyzed for the most-frequent topics and the most-referred users by assessing the usage of hashtags and mentions. The Tweets were examined by extracting the top-10 hashtags and mentions using the '*advertools 0.13.0*' module (https://pypi.org/project/advertools/). We compare the similarities and differences in the tweeting habits of health organizations and leaders of the top-10 COVID-19 resilient countries.

Computational resources and GitHub

The analysis was done using the Digital Research Alliance of Canada's service (https://www.computecanada.ca/research-portal/accessing-resources/available-resources/). The computational resources provided by the '*graham*' cluster of the Digital Research Alliance of Canada were as listed below:

1. **CPU**: 2× Intel E5-2683 v4 Broadwell @ 2.1 GHz
2. **Memory (RAM)**: 30 GB

The supplementary material for this study–data, code, and results are available on the GitHub repository (https://github.com/manmeetkaurbaxi/Societal-Impact-on-Twitter).

RESULTS

Online societal association

Prime Minister of the U.A.E, Mohammed bin Rashid Al Maktoum (online societal association: 1.000), had the highest online societal association overall, followed by Canadian Prime Minister Justin Trudeau (online societal association: 0.068) and Turkish President Recep Tayyip Erdoğan (online societal association: 0.033), among the leaders (Fig. 2A). Out of the health organizations (refer Fig. 2B), the Health Canada and Public Health Agency of Canada (PHAC, online societal association: 1.000) had the highest online societal association, followed by the Ministry of Health of the Republic of Turkey (online societal association: 0.632) and the Ministry of Health of Spain (online societal association: 0.094). The results for each of the factors affecting the online societal association, are individually explained in the following subsections.

Engagement with impact

The user impact was scaled between the range 0 and 1 (1 denotes high user impact, and 0 denotes low user impact). The results indicate that the Turkish President (Recep Tayyip

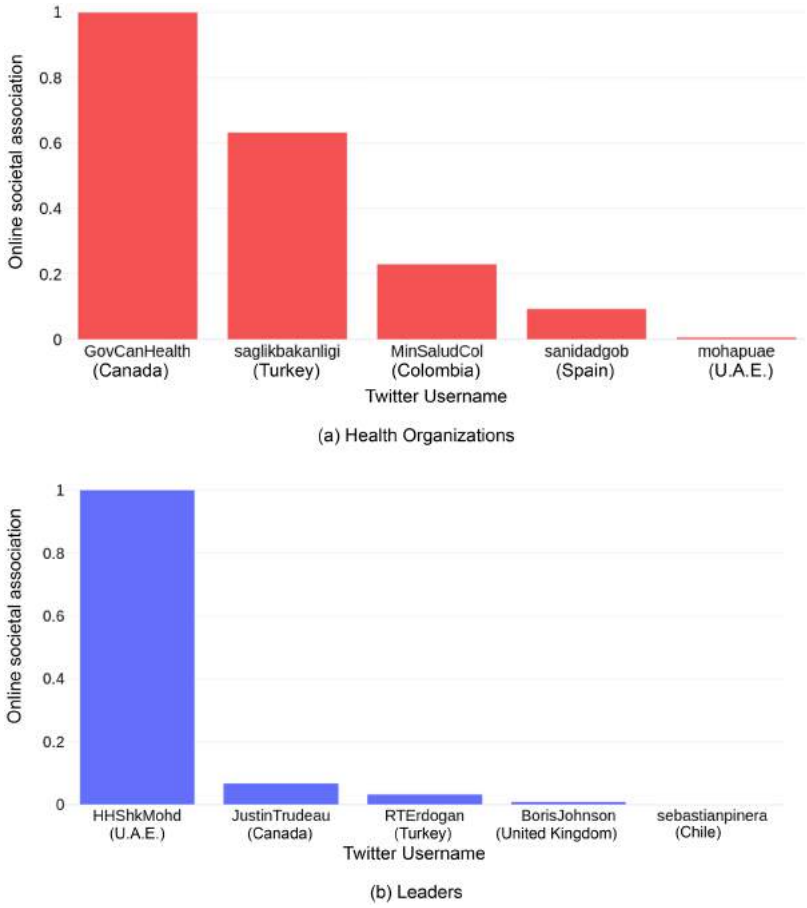


Figure 2 Online societal association (scaled) of the top-5 most impactful users by their user group, *i.e.*, (A) health organizations, and (B) leaders of the top-10 COVID-19 resilient countries.
Full-size DOI: 10.7717/peerj-cs.1121/fig-2

Erdoğan) had the greatest user impact (1.000), followed by U.K. Prime Minister (Boris Johnson, user impact: 0.978), and the Prime Minister of U.A.E (Mohammed bin Rashid Al Maktoum, user impact: 0.663) among the leaders. Among the health organizations, the Ministry of Health of the Republic of Turkey had the highest user impact (1.000), followed by the Ministry of Health and Social Protection of Colombia (user impact: 0.887) and the UK Health Security Agency (user impact: 0.778).

Among the health organizations, The Ministry of Health of the Republic of Turkey's user engagement is considerably higher than the other organizations (Fig. 3A). The highest engagement was observed during April, 2020. This can be attributed to the impacts of COVID-19, specifically, the curfew mandate imposed by the Turkish government during this time. The user engagement gradually decreased, as the COVID-19 measures lifted, and the normalization process continued. Similar to the health organizations, Turkish President Recep Tayyip Erdoğan's user engagement (as shown in Fig. 3B) is considerably higher than the other leaders, with the highest engagement recorded during August–October, 2020. The initial rise in engagement came in response to the finding of

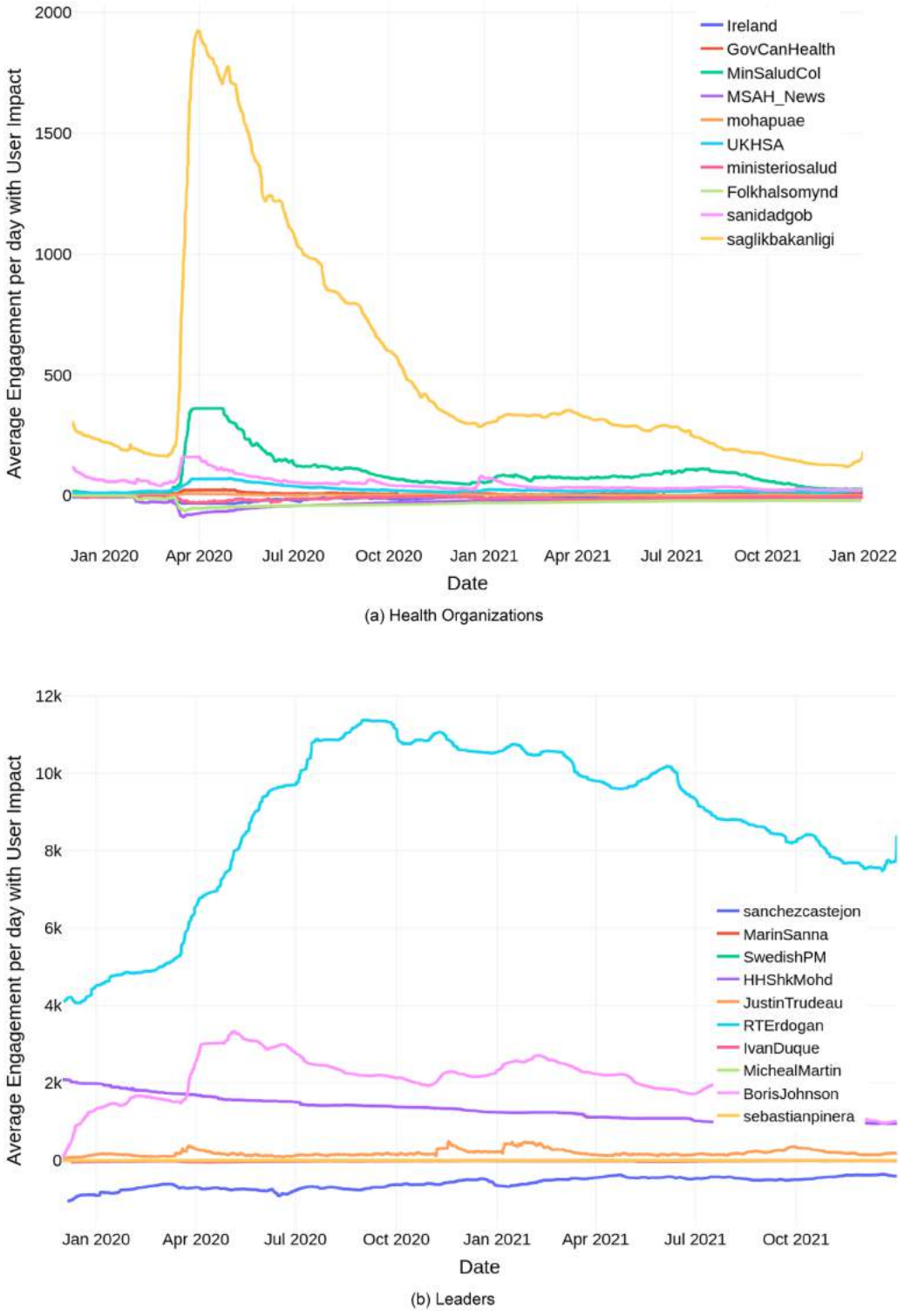


Figure 3 **Average engagement per day with user impact for (A) health organizations, and (B) leaders ofthe top-10 COVID-19 resilient countries.** Full-size DOI: 10.7717/peerj-cs.1121/fig-3

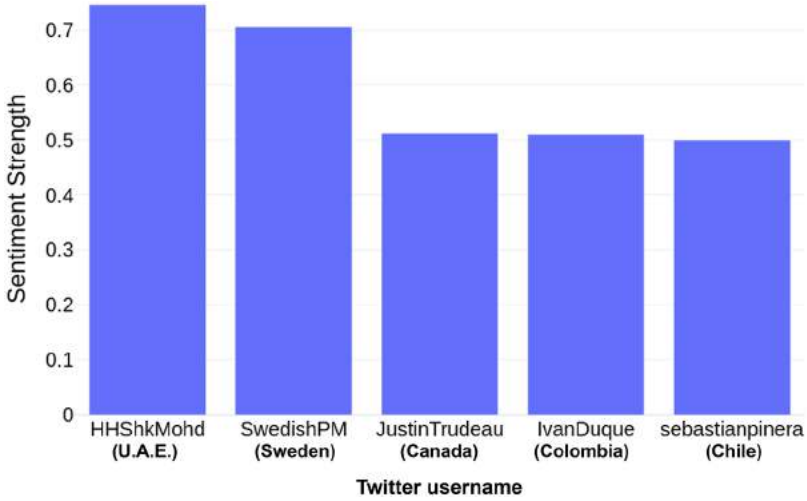


Figure 4 **Sentiment strength of top-5 leaders.** Full-size DOI: 10.7717/peerj-cs.1121/fig-4

320 billion cubic metres of natural gas in the Black Sea, which was made possible by drilling in the Danube-1 well, which began on July 20, 2020, as part of their goal of being a massive energy exporter (*Cohen, 2020*). The subsequent spike in engagement occurred in the aftermath of the 6.6 magnitude earthquake that struck Izmir, Turkey on October 30, 2020, with the government agencies rallying to save people who were trapped (*OCHA, 2020*).

Sentiment strength

After computing the Sentiment Strength for all the users, it was found that most of the users had a neutral outlook on Twitter, except for the UK Health Security Agency (UKHSA), who had a highly negative opinion (sentiment strength: −0.999). Only six user accounts out of 20 reflected positive sentiment through their Tweets; five of these were the leaders of Chile, U.A.E., Canada, Colombia, Sweden (as shown in Fig. 4) and the official account of the Public Health Agency of Canada (PHAC) (sentiment strength: 0.411). Among the leaders, U.A.E.'s Prime Minister Mohammed bin Rashid Al Maktoum had the highest positive sentiment strength (*i.e.*, 0.746), followed by the Swedish Prime Minister, Magdalena Andersson (sentiment strength: 0.706) and Canadian Prime Minister, Justin Trudeau (sentiment strength: 0.512). Figure 4 depicts the sentiment strength of the top-5 leaders.

Inclusivity and diversity strength

Leaders of the top-10 COVID-19 resilient countries were more inclusive and diverse while tweeting compared to the health organizations of these countries. Among the leaders, U.A.E's Prime Minister had the highest inclusivity and diversity strength (*i.e.*, 0.644), followed by the Colombian President, Iván Duque (inclusivity and diversity strength: 0.63), and the Chilean President, Sebastián Piñera (inclusivity and diversity strength: 0.624). For the health organizations, the results were slightly different, with Finland's Ministry of Social Affairs and Health having the highest inclusivity and diversity

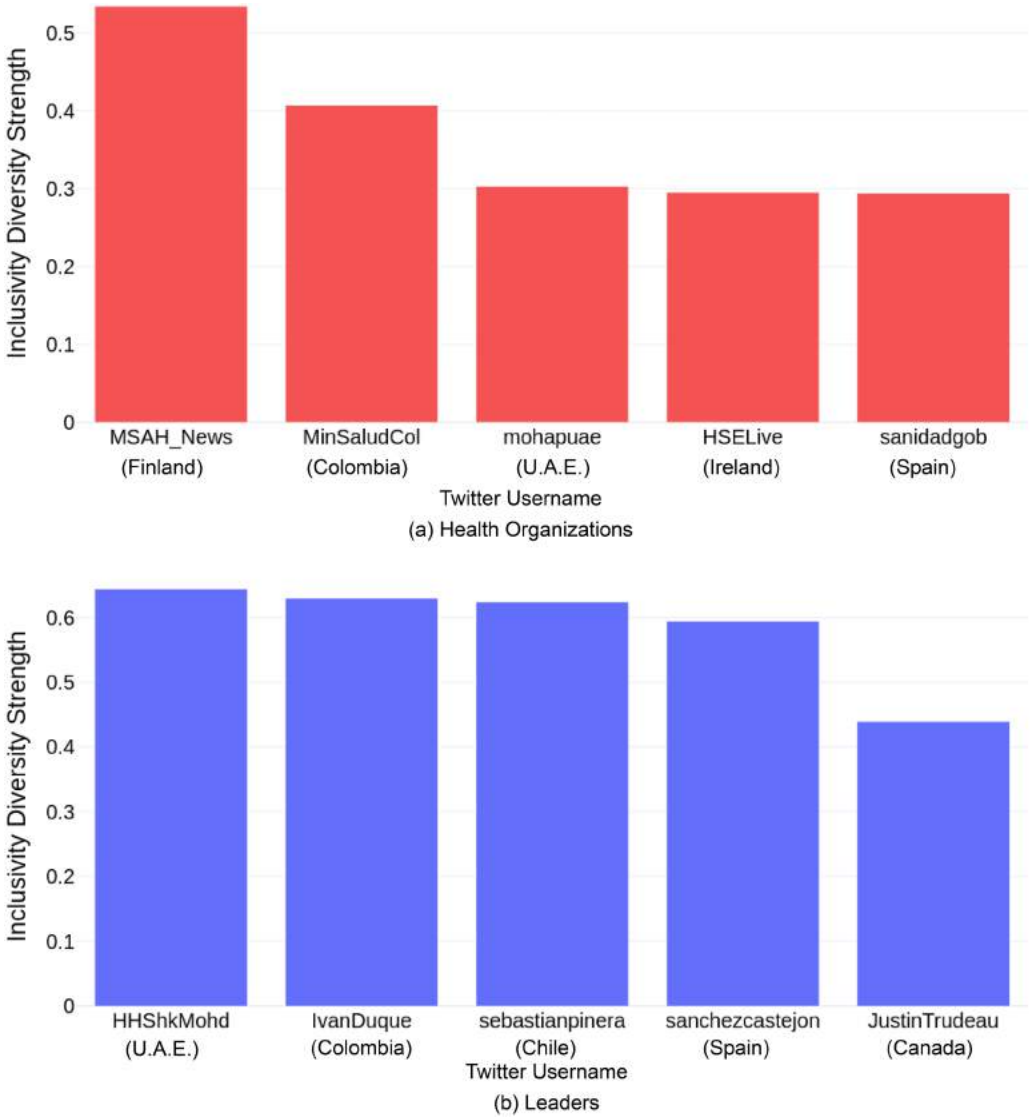


Figure 5 **Inclusivity and diversity strength for top-5 (A) health organizations, and (B) leaders.**
Full-size DOI: 10.7717/peerj-cs.1121/fig-5

strength (*i.e.*, 0.534), followed by the Colombian Ministry of Health and Social Protection (inclusivity and diversity strength: 0.407) and U.A.E.'s health organization, MOHAP (inclusivity and diversity strength: 0.303). Figures 5A and 5B illustrates the *Inclusivity and Diversity Strength* of the top-5 health organizations and leaders respectively.

Content analysis

The analysis of the hashtags served as a representation of the content, for each of the users. '*COVID-19*' was the most discussed topic among the accounts of the health organizations-as shown in Fig. 6A which displays the high frequency of '*#covid19*', '*#covid_19*', '*#yomevacuno*' (referring to the vaccination plans and status of Chile) (https://www.gob.cl/yomevacuno/), and '*#coronavirus*' hashtags in user's Tweets. The data

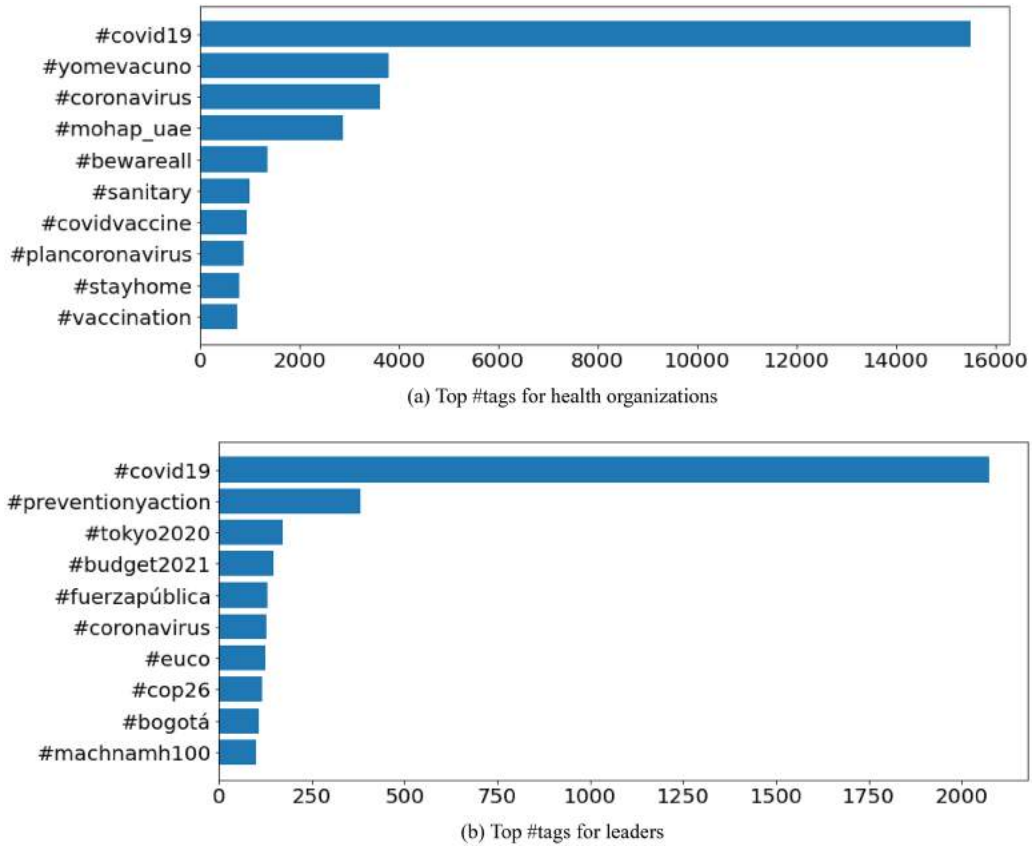


Figure 6 Top #tags for (A) health organizations, and (B) leaders of the top-10 COVID-19 resilient countries. Full-size DOI: 10.7717/peerj-cs.1121/fig-6

indicates that the health organizations communicated about the COVID-19 pandemic with the same, or similar words. However, the results of the political leaders indicated contrasting content discussion. This is due to the fact that the political leaders would discuss relevant political issues, respective to their country-attributing to the diversity of hashtags. This concept is supported by Fig. 6B which shows the variety of hashtags related to different topics; '*#covid19*', '*#cop25*' (referencing the 25th United Nations Climate Change Conference, held from December 2 to 13, 2019 (https://unfccc.int/cop25)), '*#tokyo2020*', '*#budget2021*', '*#euco*' (regarding the European Council (https://www.consilium.europa.eu/en/european-council/)), '*#Bogotá*' (referring to the capital of Columbia), '*#Fuerzaública*' (referencing the public force of Columbia (https://www.constitucioncolombia.com/titulo-7/capitulo-7)), and '*#machnamh100*' (regarding an initiative of Ireland's President, Michael D. Higgins, which explores influential events during Ireland's Decade of Commemorations (https://president.ie/en/news/article/machnamh-100-president-of-irelands-centenary-reflections)). The frequency of mentions were measured to understand the interactions of each account. Figure 7A indicates the frequency of mentions among the health organizations. The mentioned accounts are current or former health ministers, or politicians, respective to each of the top-10 COVID-19 resilient countries selected for our analysis. Figure 7B represents the most recurrent

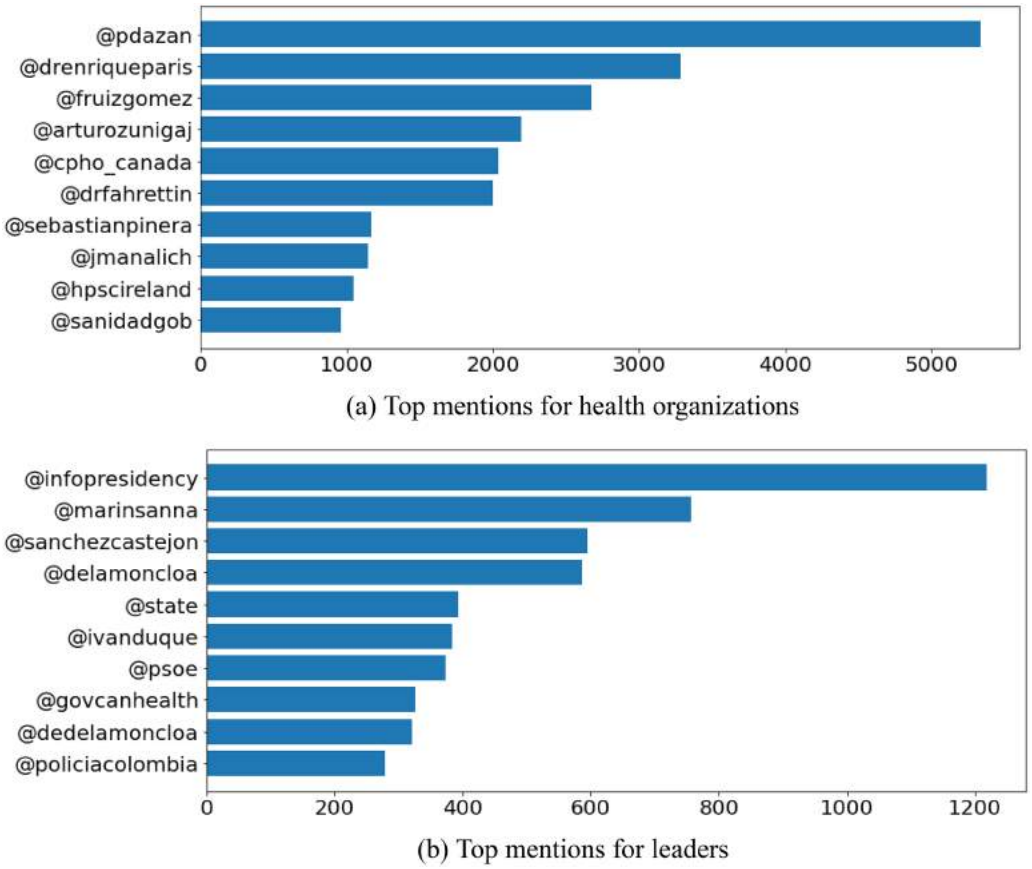


Figure 7 Top mentions for (A) health organizations, and (B) leaders of the top-10 COVID-19 resilient countries. Full-size DOI: 10.7717/peerj-cs.1121/fig-7

mentions for the political leaders. The mentioned accounts were relevant political figures or organizations to the country.

DISCUSSION AND CONCLUSION

Principal findings

Through this research, we proposed a framework for extensive analysis of social media content and the online societal associations of healthcare organizations and leaders of the top-10 COVID-19 resilient nations using NLP-based text-mining and statistical methodologies. We evaluated reasonably significant amounts of textual data for assessing the online societal association by analyzing impact and engagement, sentiment strength, and inclusion and diversity strength. The significant conclusions from our research are as follows:

- Being the most active user on social media does not necessarily imply a higher level of online societal association. The Prime Ministers of the United Arab Emirates and Canada had significantly more online societal association than the leaders of Colombia and Spain, despite the latter's having a higher number of Tweets. A similar observation is made for the health organizations, where the Canadian and Turkish health agencies

created a substantially more significant online societal association than Colombia and Ireland. According to our findings, people are also more inclined to engage with neutral Tweets, which normally contain some sort of public notification, rather than entirely positive or negative Tweets. This might imply that healthcare organizations and leaders can use this information to their advantage when developing content for social media postings to maximize their online societal associations.

- Using specific hashtags undoubtedly aids in driving engagement, as we have seen that most public engagement is highly slanted towards Tweets containing hashtags related to '*COVID-19*'. Furthermore, we note that user engagement for both the user groups, *i.e.*, health organizations and leaders, follows a predictable pattern, with peaks emerging around events of emergency or public welfare announcements (For instance, there was more public interaction when the Turkish President announced the discovery of 320 billion cubic metres of natural gas in the Black Sea during August–October 2020 and when the Ministry of Health of the Republic of Turkey tweeted about the curfew mandate in April 2020).
- Leaders of the top-10 COVID-19 resilient nations targeted wider audiences than their health organizations when it came to inclusion and diversity. Additionally, they portrayed a comparably higher sentiment strength from the health organizations.

Limitations and future work

The results of this study are confined to the COVID-19 timeline selected, *i.e.*, between December 1, 2019, to December 31, 2021. To further comprehend the online societal association of leaders and health organizations in different timeframes, the researchers might use alternative approaches to organize their data. Moreover, our research focuses on leaders' and healthcare organizations' Twitter data, which is often clean and requires little pre-processing. Because our research was confined to textual data, we could not account for the influence of image characteristics or knowledge graphs related to individual Tweets. However, it would be intriguing to see how this methodology behaves on the Tweets of other cabinet members and decision-makers of these countries, as well as investigate the organic and paid audiences, if any exist. Another area that future research might look at is the demographics of the individuals who are interacting with these contents and the penetration of Twitter among different countries.

CONCLUSION

This study investigated the online activity of healthcare organizations and leaders of the top 10 COVID-19 resilient nations on Twitter. The NLP-based statistical methods analysis of the social media activity presented here can be utilized to gauge the online societal association on the previously published Tweets and to generate Tweets that create an impact on people accessing healthcare information *via* SMPs. Each individual characteristic, *i.e.*, public impact and involvement, sentiment strength, inclusivity and diversity strength, played an equally important role in determining a user's online societal association. Thus, we believe that quantifying the online societal association and analyzing

the Tweet content provides a better understanding of how posting the appropriate Tweet at the right time may make all the difference in society.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Digital Research Alliance of Canada and CASES building, Lakehead University for providing the computational resources needed to complete this research, and DaTALab member Aditya Singhal for proof-reading the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Manmeet Kaur Baxi is funded by an AI Scholarship from Vector Institute, Toronto; an NSERC Discovery Grant (RGPIN-2017-05377) is held by Vijay Mago. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Vector Institute, Toronto.
NSERC Discovery: RGPIN-2017-05377.

Competing Interests

Vijay Mago is an Associate Editor for IEEE Access and BMC Medical Informatics and Decision Making, and an Academic Editor for PeerJ.

Author Contributions

- Manmeet Kaur Baxi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Joshua Philip conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Vijay Mago conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code and data are available at GitHub: https://github.com/manmeetkaurbaxi/Societal-Impact-on-Twitter.

REFERENCES

Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. 2020. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating

effects of public health awareness and behavioral changes: integrated model. *Journal of Medical Internet Research* **22(8)**:e19996 DOI 10.2196/19996.

Benetoli A, Chen T, Aslani P. 2018. How patients' use of social media impacts their interactions with healthcare professionals. *Patient Education and Counseling* **101(3)**:439–444 DOI 10.1016/j.pec.2017.08.015.

Bertot JC, Jaeger PT, Munson S, Glaisyer T. 2010. Social media technology and government transparency. *Computer* **43(11)**:53–59 DOI 10.1109/MC.2010.325.

Bhat V, Yadav A, Yadav S, Chandrasekaran D, Mago V. 2021. AdCOFE: advanced contextual feature extraction in conversations for emotion classification. *PeerJ Computer Science* **7(4)**:e786 DOI 10.7717/peerj-cs.786.

Bhattacharya S, Srinivasan P, Polgreen P. 2017. Social media engagement analysis of us federal health agencies on Facebook. *BMC Medical Informatics and Decision Making* **17(1)**:1–12 DOI 10.1186/s12911-017-0447-z.

Bonsón E, Bednárová M. 2018. The use of YouTube in western European municipalities. *Government Information Quarterly* **35(2)**:223–232 DOI 10.1016/j.giq.2018.04.001.

Bonsón E, Perea D, Bednárová M. 2019. Twitter as a tool for citizen engagement: an empirical study of the Andalusian municipalities. *Government Information Quarterly* **36(3)**:480–489 DOI 10.1016/j.giq.2019.03.001.

Bonsón E, Ratkai M. 2013. A set of metrics to assess stakeholder engagement and social legitimacy on a corporate Facebook page. *Online Information Review* **37(5)**:787–803 DOI 10.1108/OIR-03-2012-0054.

Bonsón E, Royo S, Ratkai M. 2015. Citizens' engagement on local governments' Facebook sites. An empirical analysis: the impact of different media and content types in Western Europe. *Government Information Quarterly* **32(1)**:52–62 DOI 10.1016/j.giq.2014.11.001.

Bonsón E, Royo S, Ratkai M. 2017. Facebook practices in Western European municipalities: an empirical analysis of activity and citizens' engagement. *Administration & Society* **49(3)**:320–347 DOI 10.1177/0095399714544945.

Bonsón E, Torres L, Royo S, Flores F. 2012. Local e-government 2.0: social media and corporate transparency in municipalities. *Government Information Quarterly* **29(2)**:123–132 DOI 10.1016/j.giq.2011.10.001.

Borah P. 2016. Political Facebook use: campaign strategies used in 2008 and 2012 presidential elections. *Journal of Information Technology & Politics* **13(4)**:326–338 DOI 10.1080/19331681.2016.1163519.

Bruns A, Moon B. 2018. Social media in Australian federal elections: comparing the 2013 and 2016 campaigns. *Journalism & Mass Communication Quarterly* **95(2)**:425–448 DOI 10.1177/1077699018766505.

Chun S, Shulman S, Sandoval R, Hovy E. 2010. Government 2.0: making connections between citizens, data and government. *Information Polity* **15(1, 2)**:1–9 DOI 10.3233/IP-2010-0205.

Cohen A. 2020. Turkey finds enormous gas field in the Black Sea—but Tricky process ahead. Forbes. *Available at https://www.forbes.com/sites/arielcohen/2020/09/18/turkeys-new-natural-gas-find-in-the-black-sea-exciting-but-tricky-process-ahead/?sh=7e437a485a86*.

Dang-Xuan L, Stieglitz S, Wladarsch J, Neuberger C. 2013. An investigation of influentials and the role of sentiment in political communication on Twitter during election periods. *Information, Communication & Society* **16(5)**:795–825 DOI 10.1080/1369118X.2013.783608.

Garcia K, Berton L. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* **101(1)**:107057 DOI 10.1016/j.asoc.2020.107057.

Gates B. 2020. Responding to COVID-19—a once-in-a-century pandemic? *New England Journal of Medicine* **382(18)**:1677–1679 DOI 10.1056/NEJMp2003762.

Gruzd A, Lannigan J, Quigley K. 2018. Examining government cross-platform engagement in social media: Instagram vs Twitter and the big lift project. *Government Information Quarterly* **35(4)**:579–587 DOI 10.1016/j.giq.2018.09.005.

Gruzd A, Roy J. 2016. Social media and local government in Canada: an examination of presence and purpose. In: *Social Media and Local Governments*. Berlin: Springer, 79–94.

Haman M. 2020. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. *Heliyon* **6(11)**:e05540 DOI 10.1016/j.heliyon.2020.e05540.

Hollebeek LD, Glynn MS, Brodie RJ. 2014. Consumer brand engagement in social media: conceptualization, scale development and validation. *Journal of Interactive Marketing* **28(2)**:149–165 DOI 10.1016/j.intmar.2013.12.002.

Househ M. 2013. The use of social media in healthcare: organizational, clinical, and patient perspectives. *Enabling Health and Healthcare Through ICT: Available, Tailored and Closer* **183**:244–248 DOI 10.3233/978-1-61499-203-5-244.

Jünger J, Fähnrich B. 2020. Does really no one care? Analyzing the public engagement of communication scientists on Twitter. *New Media & Society* **22(3)**:387–408 DOI 10.1177/1461444819863413.

Li H, Sakamoto Y. 2014. Social impacts in social media: an examination of perceived truthfulness and sharing of information. *Computers in Human Behavior* **41(1)**:278–287 DOI 10.1016/j.chb.2014.08.009.

Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. 2021. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the reddit social media platform: a call to action for strengthening vaccine confidence. *Journal of Infection and Public Health* **14(10)**:1505–1512 DOI 10.1016/j.jiph.2021.08.010.

Merchant RM, Lurie N. 2020. Social media and emergency preparedness in response to novel coronavirus. *JAMA* **323(20)**:2011–2012 DOI 10.1001/jama.2020.4469.

Norris DF, Reddick CG. 2013. Local e-government in the united states: transformation or incremental change? *Public Administration Review* **73(1)**:165–175 DOI 10.1111/j.1540-6210.2012.02647.x.

OCHA. 2020. İzmir/Turkey: Earthquake Situation Report No: 01, 30 October 2020. *Available at https://reliefweb.int/report/turkey/zmirturkey-earthquake-situation-report-no-01-30-october-2020*.

Parmelee JH, Bichard SL. 2011. *Politics and the Twitter revolution: how Tweets influence the relationship between political leaders and the public*. Lanham: Lexington books.

Patel KD, Heppner A, Srivastava G, Mago V. 2019. Analyzing use of Twitter by diabetes online community. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Piscataway: IEEE, 937–944.

Patel KD, Zainab K, Heppner A, Srivastava G, Mago V. 2020. Using Twitter for diabetes community analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics* **9(1)**:1–16 DOI 10.1007/s13721-020-00241-y.

Peck JL. 2014. Social media in nursing education: responsible integration for meaningful use. *Journal of Nursing Education* **53(3)**:164–169 DOI 10.3928/01484834-20140219-03.

Qudar MMA, Bhatia P, Mago V. 2021. Onset: opinion and aspect extraction system from unlabelled data. In: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Piscataway: IEEE, 733–738.

Ríos A-M, Benito B, Bastida F. 2017. Factors explaining public participation in the central government budget process. *Australian Journal of Public Administration* **76(1)**:48–64 DOI 10.1111/1467-8500.12197.

Rufai SR, Bunce C. 2020. World leaders' usage of twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health* **42(3)**:510–516 DOI 10.1093/pubmed/fdaa049.

Russmann U, Svensson J. 2017. Interaction on Instagram?: glimpses from the 2014 Swedish elections. *International Journal of E-Politics* **8(1)**:50–66 DOI 10.4018/IJEP.

Sahly A, Shao C, Kwon KH. 2019. Social media for political campaigns: an examination of Trump's and Clinton's frame building and its effect on audience engagement. *Social Media & Society* **5(2)**:2056305119855141 DOI 10.1177/2056305119855141.

Sampieri RH. 2018. *Metodologa de la investigación: las rutas cuantitativa, cualitativa y mixta*. México: McGraw Hill.

Sandoval-Almazan R, Valle-Cruz D. 2018. Facebook impact and sentiment analysis on political campaigns. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. 1–7.

Shah N, Srivastava G, Savage DW, Mago V. 2020. Assessing Canadians health activity and nutritional habits through social media. *Frontiers in Public Health* **7**:400 DOI 10.3389/fpubh.2019.00400.

Siebers V, Gradus R, Grotens R. 2019. Citizen engagement and trust: a study among citizen panel members in three Dutch municipalities. *The Social Science Journal* **56(4)**:545–554 DOI 10.1016/j.soscij.2018.09.010.

Singhal A, Baxi MK, Mago V. 2022. Synergy between public and private healthcare organizations during COVID-19 on Twitter. *JMIR Medical Informatics* **10(8)**:e37829 DOI 10.2196/37829.

Tiedemann J. 2020. The tatoeba translation challenge-realistic data sets for low resource and multilingual mt. *ArXiv preprint*. DOI 10.48550/arXiv.2010.06354.

Ventola CL. 2014. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics* **39(7)**:491.

Vesnic-Alujevic L, Van Bauwel S. 2014. YouTube: a political advertising tool? A case study of the use of YouTube in the campaign for the European parliament elections. *Journal of Political Marketing* **13(3)**:195–212 DOI 10.1080/15377857.2014.929886.

PeerJ Computer Science

Effect of visual imagery in COVID-19 social media posts on users' perception

Waleed M. Al-nuwaiser

Department of Computer Science, College of Computer and Information Sciences, Al-Imam Mohamed Ibn Saud Islamic University, Riyadh, Saudi Arabia

ABSTRACT

People receive a wide variety of news from social media. They especially look for information on social media in times of crisis with the desire to assess the risk they face. This risk assessment, and other aspects of user reactions, may be affected by characteristics of the social media post relaying certain information. Thus, it is critical to understand these characteristics to deliver information with the reactions in mind. This study investigated various types of imagery used as thumbnails in social media posts regarding news about the COVID-19 pandemic. In an experimental design, 300 participants viewed social media posts containing each of the three types of imagery: data visualization (directly about risk information), advisory (not containing direct risk information, but instead help on how to lower risk), or clickbait (containing no risk-related information, just generic visuals). After observing the social media posts, they answered questionnaires measuring their emotions (valence, arousal, and dominance), risk perception, perceived credibility of the post, and engagement. The participants also indicated their emotions towards and interest in COVID-19 news coverage, age, gender, and how often and actively they use social media. These variables acted as controls. The data were analysed using mixed linear models. Results indicated that advisory imagery positively influenced valence, arousal, dominance, credibility, and (lower) risk perception. Alternatively, imagery showing data visualizations yielded low valence, arousal, dominance, credibility, and high risk perception. Clickbait-styled thumbnails which carry no information are usually measured between the other two types. The type of imagery did not affect the motivation to engage with a post. Aside from visual imagery, most variables were affected by COVID sentiment and the usual activity on social media. These study results indicate that one should use advisory imagery for more comfortable news delivery and data visualization when the poster wishes to warn users of existing risks.

Subjects Bioinformatics, Human-Computer Interaction, Brain-Computer Interface
Keywords Imagery posts, Social media, COVID-19, SAM, Valence, Arousal, Dominance, Risk perception, Credibility and engagement

Submitted 5 September 2022
Accepted 18 October 2022
Published 15 November 2022

Corresponding author
Waleed M. Al-nuwaiser,
wmalnuwaiser@imamu.edu.sa

Academic editor
Jude Hemanth

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.1153**

Copyright
2022 Al-nuwaiser

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Social media have become prominent sources of information for many people (*Westerman, Spence & Van Der Heide, 2014*). Therefore, it was a matter of time before social media started being used as a tool for swaying public opinion through fake news (*Deutsch, 2017*; *English, 2017*; *Ghanem, Rosso & Rangel, 2020*; *Grech & Masukume, 2017*; *Guess, Nagler & Tucker, 2019*) and even trying to win over elections (*Allcott & Gentzkow,*

2017; Barthel, Mitchell & Holcomb, 2016; Shane, 2017). While these are bad influences, social media may exert good influences as well—it provides rapid communication in times of crisis, such as during the cholera outbreak in Haiti (Sutter, 2010). Therefore, it is crucial to investigate how social media posts may be framed and shown to users to show credible information and have them see it as credible.

BACKGROUND

The three-component emotion model

Ever since the pioneering work of Wundt (1913), there has been a line of thought in psychology that emotions can be assessed through three dimensions. Wundt termed them *lust* (pleasure), *spannung* (tension), and *beruhigung* (inhibition). In later years, other scholars determined that three dimensions precisely describe emotions, and although they used slightly different names, they referred to very similar phenomena. For instance, a group of researchers (Osgood, Suci & Tannenbaum, 1957) utilized factor analysis on participants' estimations of words on 50 bipolar scales (for example, hot or cold). They determined that three factors can successfully explain them—evaluation, activity, and potency. They have also shown that these aspects of emotion are usable in different cultures and for non-verbal stimuli (Osgood, Suci & Tannenbaum, 1957). Further research found that the three dimensions assess body parts, facial gestures (Mehrabian, 1970), and human interactions (Mehrabian & Russell, 1974). Henceforth, the semantic differential scale developed by Mehrabian & Russell (1974) was widely used to measure the three dimensions of emotion. It contains 18 pairs of opposite adjectives (for instance, happy—unhappy), and between the adjectives, there is a nine-point scale (ranging from −4 to 4), on which the participant can indicate how close they are to either of the poles. Then, they calculate a composite score to measure the degree to which each of the three dimensions of emotion is present in a participant. This methodology was and is still widely used, but there was also room for improvement. If many stimuli need to be measured, the test can be very time-consuming, and it is language-dependent.

Those shortcomings are why Prof. Peter Lang and his colleagues developed the self-assessment mannequin (SAM) method (Bradley & Lang, 1994; Lang, 1980). This method, which will be described in more detail later, only utilizes three scales and measures the same three dimensions of emotion, albeit named slightly differently—valence, arousal, and dominance. The valence dimension relates to how positive or negative an emotion is. The arousal dimension relates to how excited or calm the person is. Finally, the dominance dimension relates to how much in control the person feels, compared to feeling like they do not influence the situation. As will be shown more extensively, researchers have successfully used this methodology to measure these dimensions of emotion in a fast and exact manner in various contexts (Isomursu et al., 2007).

Social effects of the COVID-19 pandemic

The emergence of the COVID-19 pandemic has had incredibly destructive effects on people worldwide, both in terms of physical illness and death and regarding psychological and social effects. As any other crisis does, the pandemic changed how people lived. They

stayed at home much more, made less contact with their friends and loved ones, and changed how they worked or studied (through online media). This behaviour brought about many issues with mental health (*AlAzzam et al., 2021*; *Almandoz et al., 2021*; *Biro-Hannah, 2021*; *Cao et al., 2020*; *Liang et al., 2020*; *Pietrabissa et al., 2021*; *Vahedian-Azimi et al., 2020*) as societal changes influenced people.

To alleviate these changes, people commonly turned to social media networks to connect with others and find information, which is typical for crises. For example, the Zika virus outbreak has caused people to use social media for information and assurance. More specifically, a study (*Gui et al., 2017*) demonstrated that social media was the primary channel people used to assess the risk associated with the Zika virus outbreak in 2016. Similarly, the ongoing COVID-19 pandemic inspired a lot of news coverage, fear, and risk in everyday lives. Analyses of social media post sentiments showed that, during the beginning of the pandemic, people were experiencing multiple pandemic-related threats, including illness, secondary health conditions, economic, socio-behavioural, and institutional risks. They also used social media to find information which would help them with their day-to-day decision making (*Pine et al., 2021*). Therefore, it is crucial to understand how different ways of shaping information on social media may influence users' perceptions of and reactions to such information.

RELATED WORK

In recent years, research has explored various aspects of social media experiences. A critical aspect of reading information on social media is knowing who posted it and deliberating on whether that source is trustworthy. To understand the factors which contribute to (not) trusting a source, *Kim & Dennis (2019)* conducted an experimental study. They manipulated the alleged source of an article and tested whether it nudged the readers towards taking a closer look at the source. The result of the nudging was that the users were more sceptical when attention was brought to the source but were more trusting towards the article (rated it with higher believability) when the source had a high rating. Increased believability encouraged higher engagement with the article. Researchers measured this engagement through interaction with the post—liking, sharing, or commenting upon it. In the remainder of this article, engagement means interacting with a post. They also had more trust in articles that aligned with their views, thus manifesting confirmation bias (*Kim, Moravec & Dennis, 2019*; *Kim & Dennis, 2019*).

As mentioned earlier, social media is a vital source of information during crises (*Gui et al., 2017*; *Pine et al., 2021*). One of the main types of information people seek during an emergency is risk information—How much in danger are we, and should we be scared? Throughout the COVID-19 pandemic, especially at the beginning, people were frightened and saw the risk of dangerous health outcomes happening to them and everyone else as very high (*Dryhurst et al., 2020*). People both expressed anxiety and sought answers and knowledge online, which would help them deal with the new situation, as shown in a study on the Korean social media space at the beginning of the pandemic (*Jo et al., 2020*). As already mentioned, a large *corpus* of research (*e.g.*, *Bica et al., 2019*; *Gui et al., 2017*; *Pine et al., 2021*) shows that risk assessment occurs primarily through social media, which

means that health officials should strive to utilize social media to communicate risk precisely. They should do it knowing how every element of a social media post affects the potential reader.

However, only a handful of studies have investigated the effects that various aspects of social media posts have on users. One line of research by *Kim, Moravec & Dennis (2019)* focused on fake news and how reliability ratings can push users towards being more sceptical. However, none of the mainstream social media platforms (Facebook, Instagram, Twitter, YouTube, and TikTok) have credibility ratings. Another line of research is investigating how headlines used in social media posts affect user perceptions. In that vein, *Janét, Richards & Landrum (2022)* investigated three types of headlines about environmental topics: traditional ones (with straightforward information), those that ask questions, and forward-referencing headlines ('You'll never believe what…'). They found that headlines did not affect story engagement or selection but did affect credibility, and the question-formatted headlines seemed the least credible. The participants' political views and scientific curiosity also affected the engagement and credibility of the stories (*Janét, Richards & Landrum, 2022*). Another study found that users typically share news on social media through more interpersonal and subjective methods. Recent studies have called for more investigation into how they share news and which determinants lead to which outcomes in their perception (*Welbers & Opgenhaffen, 2019*).

Another study aimed to answer this question investigated how the emotional tone of a news social media post affects users' engagement. *Choi, Lee & Ji (2021)* examined ~15,000 social media posts using software to recognize emotions in the thumbnails and the text. They determined that social media posts with negative text were more commented on and shared, while positive texts earned more reactions. Furthermore, sadness was the thumbnail emotion that yielded the most engagement. Other researchers have shown that visual aspects of a social media post can also be important in relaying risk-related information regarding natural disasters (*Bica et al., 2019*).

Aside from the text in the post, the visual aspects of social media posts influence users as well. Also, social media is a vital space to disseminate information in times of crisis in general and during the COVID-19 pandemic specifically. Thus, this study aimed to investigate the effect of various types of imagery used in COVID-related social media posts. The study's primary research question is to determine the effects that different types of visual stimuli used in thumbnails have on how social media users perceive social media posts regarding information about COVID-19. The three types of visual stimuli are (a) those containing charts with information about the pandemic to convey risk information; (b) those with advice, which do not provide any information about the risk of the pandemic but instead just advise on how to reduce personal and community risks; and (c) those with no information and just striking visuals that help the user identify the topic of the article.

There are six specific research questions:

RQ1: What is the effect of the three types of visual stimuli on the valence of emotions that users feel after looking at a COVID-related social media post?

Table 1 Absolute and relative frequency of participants from various countries.

Country	Frequency	Percent
United States	217	72.33
India	57	19
Saudi Arabia	11	3.67
Brazil	3	1
Romania	3	1
Germany	2	0.67
Italy	2	0.67
Singapore	1	0.33
United Kingdom	1	0.33
Missing	3	1
Total	300	100

RQ2: What is the effect of the three types of visual stimuli on the arousal that users feel after looking at a COVID-related social media post?

RQ3: What is the effect of the three types of visual stimuli on the dominance that users feel after looking at a COVID-related social media post?

RQ4: What is the effect of the three types of visual stimuli on the users' perception of risk after looking at a COVID-related social media post?

RQ5: What is the effect of the three types of visual stimuli on the users' engagement with a COVID-related social media post?

RQ6: What is the effect of the three types of visual stimuli on the users' perception of the credibility of a COVID-related social media post?

Aside from the primary independent variable (type of visual stimuli), the study will record additional variables to serve as controls. These variables include the confirmation bias measure (due to the finding of *Kim, Moravec & Dennis (2019)* that it has effects on post-perception), age, and social media usage (both through time and activity).

METHODS

Participants

Three hundred participants (40.7% female) completed the survey, collected through Amazon's Mechanical Turk (MTurk) platform. Each participant in the study received a payment of $1.5. The average age of participants was 37.09 (ranging from 21 to 78, SD = 9.896). The country of the participants appears in Table 1. Participants were sampled using convenience sampling, as this was the only way to collect participants online during the pandemic within the resources available to the researcher.

Design

An experimental design measured the effect that viewing each type of imagery has on participants. Before participating in the questionnaire, participants signed an online consent form. All participants looked at posts containing three types of visual stimuli: advisory, clickbait, or data visualization. Then, they gave estimates on their emotional

Table 2 Headline sets used in each of the parallel forms of the surveys.

Type of imagery	First parallel form	Second parallel form	Third parallel form
Generic	Set 1	Set 3	Set 2
Advisory	Set 2	Set 1	Set 3
Data visualization	Set 3	Set 2	Set 1

state, perception of risk, engagement with the post, and its credibility. The creation of three parallel survey formats avoided confounding the effect of visuals with the headline content. The headlines were counterbalanced across three parallel survey forms by first choosing nine headlines and labelling them randomly from 1 to 9. Then, a random number generator created a random sequence of numbers 1 to 9. The design used the headlines corresponding to the first three numbers in the sequence for the three stimuli with generic COVID-19 imagery. It used the second three headlines for the three stimuli with informative, advisory imagery. It used the final three headlines for the three stimuli with data visualizations. Thus, it created the first parallel set of stimuli. Then, the Latin square method counterbalanced the headline sets across the remaining two parallel forms. The second form swapped the places of the three sets of headlines among the three groups of visual stimuli. The first set now used the informative, advisory imagery, the second used the data visualization stimuli, and the third used the generic COVID-19 imagery. The creation of the third parallel form utilized the same logic, with an additional step of swapping places. The first set of headlines now used the data visualization stimuli, the second used generic COVID-19 imagery, and the third used informative advisory imagery. This method ensured that the registered effect is from the types of imagery and not from the headlines. The process of counterbalancing appears in Table 2. As the table illustrates, all the headline sets appear together with each type of imagery exactly once.

MATERIALS

Social media posts

When creating fake social media posts, there were two elements to consider: imagery and headlines. The imagery came from the following sources: advice imagery from the BBC's webpage titled 'Coronavirus information: Four posters' (*BBC, 2020a*); data visualisations from the BBC's webpage titled COVID: 12 charts on how COVID changed our lives (*BBC, 2020b*); and generic Coronavirus imagery (clickbait) from various articles posted by BBC. The headlines were from the BBC's website to keep them as descriptive as possible without mention of any famous people. Examples of all three types of imagery appear in Fig. 1. The social media pages that posted the content were completely made-up, and only reflected the idea that they were posted by a news website, as has been done in previous studies (*Kim, Moravec & Dennis, 2019*). Although it is possible that the news website has an effect on credibility in the eyes of the participants, it was expected that all websites would have a similar effect. Furthermore, they were counterbalanced in the same manner as the titles (title-poster pairs were not switched when imagery was switched for the three parallel

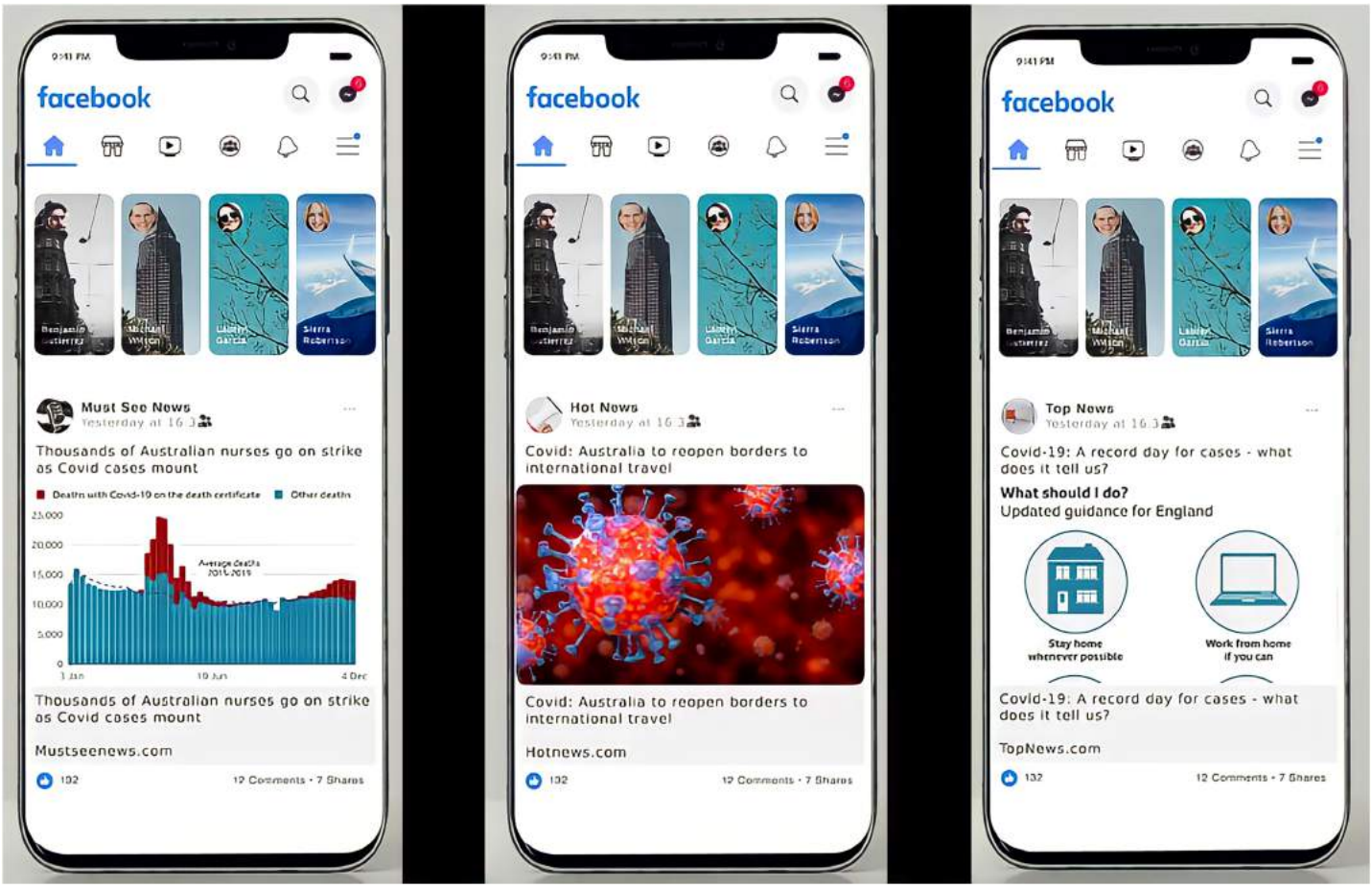


Figure 1 **Examples of used imagery, including data visualization (A), clickbait (B), and advice (C).**
Full-size DOI: 10.7717/peerj-cs.1153/fig-1

studies). As a result, there was no reason to expect that the website titles would have an effect on user perceptions that had not been controlled for by the study design. Mobile social media inspired the format of the posts, as studies show that most people use social media on their mobile phones. For example, in 2019, 79% of social media usage in the US was through mobile devices (*Dixon, 2022*).

Self-assessment mannequin (SAM)

This method used the self-assessment mannequin (*Lang, 1980*) to measure participants' emotional states after looking at social media posts. It measures three dimensions of emotion—pleasure, arousal, and dominance (*Bradley & Lang, 1994*), using non-verbal estimation on a nine-point Likert-type scale. This assessment tool is widely used for quickly and reliably determining users' reactions to technology, as it is informative for the researcher and easy to use for the participants (*Dormann, 2001*, *2003*; *Isomursu et al., 2007*). The scale appears in Fig. 2.

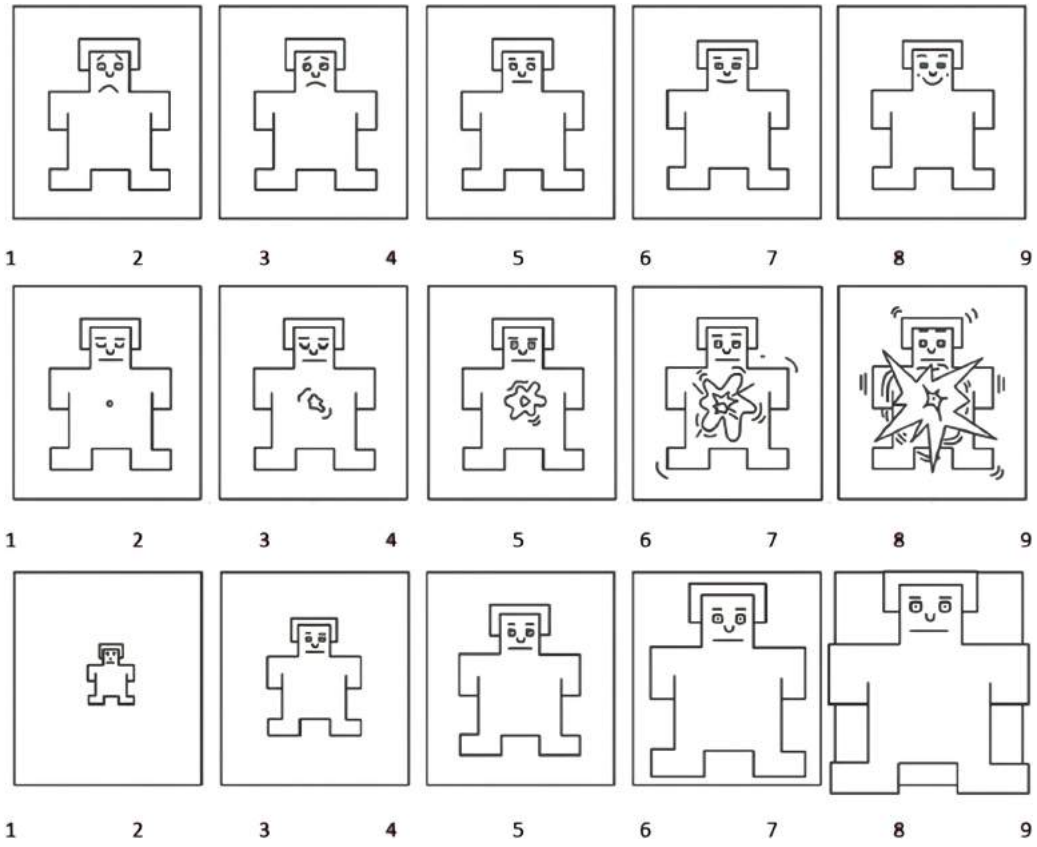


Figure 2 Self-assessment mannequin—valence, arousal, and dominance scale, in order. Figure source: © Elsevier, *Bradley & Lang (1994)*. Full-size DOI: 10.7717/peerj-cs.1153/fig-2

Risk perception

A new questionnaire measured participants' risk perception after looking at social media posts. Previous literature contained various surveys measuring multiple types of risk perception, but none of them was fully applicable to the current research context. Therefore, we created a survey with five items, answered through a five-point Likert-type (ranging from 1 to 5) scale by adapting questions from previous studies (*Alsubaie et al., 2019*; *Kim, Moravec & Dennis, 2019*; *Qin et al., 2021*). The final scale showed excellent internal reliability (α ranging from 0.89 to 0.904). An example item is 'After seeing this news post, how much anxiety do you feel about the COVID-19 pandemic?'.

Post credibility

This method adapted a previously created survey by *Janét, Richards & Landrum (2022)* to measure participant perception of the credibility of social media posts. The survey consists of ten items, answered with a six-point Likert-type scale ranging from 1 (disagree) to 6 (strongly agree). An example item is 'Based on this post, I expect that the content in the article…can be trusted.' (The part before the ellipsis is the same for all questions, and the part after it differs for each item). The scale showed good reliability both in the study of

Janét, Richards & Landrum (2022) (α = 0.88) and in the present study (α ranging from 0.781 to 0.815).

Engagement

The survey of *Janét, Richards & Landrum (2022)* measured how much the participants engaged with the post. It asked participants whether they 'would likely share the article, read the article, comment on the article, or ignore the article.' The participants selected all that applied, and the number of activities (ranging from 0 to 3) was an indicator of engagement.

COVID sentiment

A COVID sentiment index measured the participants' bias towards COVID-related social media posts. The index took inspiration from previous research on cognitive bias (*Kim, Moravec & Dennis, 2019*) by asking the participants two questions: the first question, 'What's your opinion on news' reportage on COVID-19?', was answered on a scale between −3 and 3; the second question, 'Do you find the COVID-19 pandemic important?' was answered on a scale ranging from 1 to 7. By multiplying the two answers, an index which simultaneously indicates how strongly and in what direction a participant feels about the pandemic was created. The index ranged from −21 to 21.

PROCEDURE

The researcher followed the standard ethics procedures throughout this study. Initially, participants have been thanked by the researcher for taking part in the questionnaire and informed them about the details of the study. Participants signed an online consent form. They learned that their information will not be analysed individually but only at a group level, and their data will be treated confidentially and anonymously. The researcher informed them that they were free to quit taking the survey at any point and received a contact e-mail address of the researcher if they had any inquiries. Then, they answered questions regarding their sentiment toward COVID-19 news reporting (see the materials section). Afterwards, they received a (fake) social media post, which they were instructed to view carefully. The next screen presented a question which tested whether they were paying attention to the stimulus. This question was an attention check, and participants who wrongly answered four or more questions were discarded from subsequent data analysis. There were 55 such cases, reducing the final sample to 245 participants. After the attention check, the participants answered the SAM, credibility, risk perception, and engagement surveys about the stimulus. Then, this process repeated for the following eight stimuli. After finishing all nine stimuli, they answered questions about their demographic information and social media usage. In the end, they were thanked for their participation. The whole process through which the participants went appears in Fig. 3.

RESULTS

Descriptive statistics

Brief descriptive statistics of the six main study variables appear in Table 3. Kurtosis values denote negative excess kurtosis. Valence and arousal scores were on average only slightly

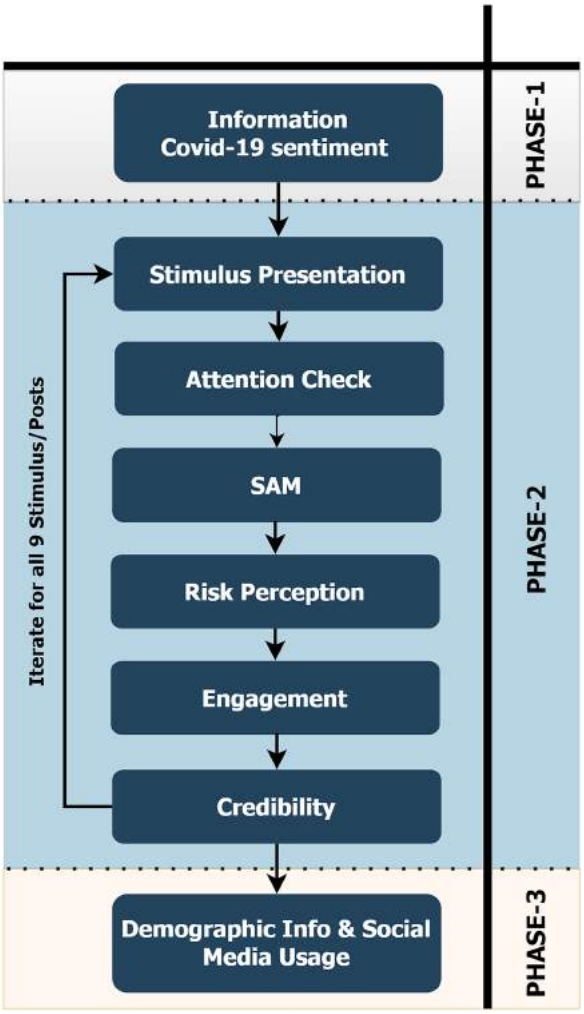


Figure 3 Diagram representing the process of taking the survey.
Full-size DOI: 10.7717/peerj-cs.1153/fig-3

Table 3 Descriptive statistics of the six main study variables.

Variable	M	SE (M)	SD	SE (SD)	Skewness (SE)	Kurtosis (SE)
Valence	5.51	0.073	1.98	0.089	−0.169 (0.090)	−0.648 (0.180)
Arousal	5.55	0.068	1.86	0.084	−0.167 (0.090)	−0.566 (0.180)
Dominance	6.00	0.072	1.97	0.089	−0.543 (0.090)	−0.282 (0.180)
Risk	3.35	0.034	0.93	0.042	−0.674 (0.090)	−0.139 (0.180)
Engagement	1.53	0.036	0.98	0.044	0.203 (0.090)	−01.066 (0.180)
Credibility	3.56	0.021	0.56	0.025	−0.597 (0.090)	1.949 (0.180)

higher than the middle value of the SAM (5). Dominance was in the same range, albeit slightly higher than the other two. The standard errors and standard deviations of all three variables were comparable. Furthermore, risk perception was slightly elevated, with the mean answer nearly a full point above the theoretical middle (2.5). It also showed low

Table 4 Cronbach's alpha across three stimuli for each stimuli type.

Variable	Clickbait	Visual	Advice
Valence	0.827	0.905	0.874
Arousal	0.856	0.869	0.877
Dominance	0.879	0.899	0.840
Risk	0.916	0.946	0.933
Engagement	0.816	0.941	0.946
Credibility	0.942	0.806	0.882

standard deviation and standard error. The mean of engagement indicates that it is exactly in line with the expected average of 1.5, with comparably slightly higher SE and SD, in comparison to risk. Finally, credibility also showed an expected mean close to the theoretical (3.5), with very low standard deviation and standard error values. The three SAM dimensions show similar means and standard deviations, and their distributions are skewed to the left and platykurtic. An exception is the dominance dimension, which is mesokurtic. Furthermore, risk perception is skewed to the left and mesokurtic, engagement is slightly skewed to the left and platykurtic, and credibility is skewed to the left and platykurtic as well. These deviations from the normal distribution agree with both Kolmogorov–Smirnoff and Shapiro–Wilk tests, which indicated significant deviations for all six variables (all $p < 0.001$).

The reliability of the three types of stimuli was determined using Cronbach's alpha coefficients. Cronbach's alpha coefficients across the three stimuli for each of the three types of stimuli were above 0.8, as shown in Table 4. This result indicates that the types of stimuli were similar enough inside the clusters of stimuli and could reliably be measured against other variables.

Results of general linear mixed models

The effects of the primary variable of interest, the type of visual material used in the thumbnails, along with the remaining predictors, including age, sentiment towards COVID-19, hours on social media, and activity on social media, will be investigated using general linear mixed models. Due to differences in some of the dependent variables based on country of origin, this variable was included in the models as well, but only as a control variable, and will not be interpreted, as this study did not aim to measure cultural differences. The researcher chose this model instead of mixed ANOVA models due to its higher robustness and customizability.

The goal of the present article was not to investigate the interactions of the visual type of thumbnails with the remaining predictors. Instead, it tests for its effect when the other variables are controlled. Therefore, the general linear mixed models were deemed more suitable. As it is unlikely that either of the between-subject variables varies inside the subjects, all of them were used as fixed effects. Furthermore, the visual type of covariance structure was determined as unstructured, as it is natural to expect high covariance among the different types of visual stimuli, and there is no reason to assume a specific pattern in

Table 5 Results of tests of fixed effects for all six criterion variables.

Predictor variables	Valence		Arousal		Dominance		Risk		Engagement		Credibility	
	F	Sig.	F	Sig.	F	Sig.	F	Sig.	F	Sig.	F	Sig.
Intercept	225.950	0.000	210.026	0.000	261.755	0.000	332.494	0.000	35.432	0.000	625.947	0.000
Visual type	19.343	0.000	4.301	0.014	14.708	0.000	11.079	0.000	0.749	0.474	12.283	0.000
Hours per day on SM	0.920	0.527	1.002	0.447	0.624	0.821	2.341	0.007	1.160	0.312	0.502	0.913
Activity on SM	11.583	0.000	16.931	0.000	11.077	0.000	18.009	0.000	17.652	0.000	3.111	0.016
Age	2.281	0.132	0.088	0.767	3.171	0.076	0.952	0.330	0.730	0.394	0.693	0.406
COVID sentiment	31.761	0.000	18.562	0.000	18.013	0.000	20.962	0.000	6.276	0.013	13.897	0.000
Country	4.858	0.028	1.496	0.222	0.013	0.909	0.586	0.445	0.090	0.764	0.067	0.0795

which that covariance will occur. The overall significant fixed effects of each of the six criterion variables appear in Table 5. Subsequent sub-sections will investigate each of these predictors in further detail.

Valence

As seen in Table 5, the valence of the reaction to the different types of visual stimuli has been successfully predicted by the type of stimulus, social media activity, and COVID sentiment. Pairwise comparisons using the Bonferroni correction showed that all pairwise differences among the visual types were significant. The advice thumbnail type had higher valence than both the bait (MD = 0.186, $p = 0.026$) and the visual (MD = 0.380, $p < 0.001$) thumbnails, and the bait had higher valence than the visual (MD = 0.380, $p < 0.001$) thumbnail. Furthermore, valence had been positively affected by activity on social media and sentiment towards COVID, indicating that participants who were more active on social media and had a more positive sentiment towards the pandemic reporting also experienced higher valence while observing the social media posts.

Arousal

As Table 5 shows, the arousal of the reaction to the different types of visual stimuli has been successfully predicted by the type of stimulus, social media activity, and COVID sentiment. Pairwise comparisons using the Bonferroni correction showed that advice thumbnails yielded higher arousal than both bait (MD = 0.171, $p = 0.035$) and visual (MD = 0.192, $p = 0.024$) types. There were no differences between bait and visual types of visuals ($p = 1$). Arousal was positively affected by activity on social media as well as sentiment towards COVID. This result indicates that participants who were more active on social media and had a more positive sentiment towards the pandemic reporting also experienced higher arousal while observing the social media posts.

Dominance

As Table 5 shows, the experienced dominance upon viewing the different types of visual stimuli has been successfully predicted by the type of stimulus, social media activity, and COVID sentiment. Participants experienced higher dominance upon viewing the advice thumbnail, in comparison to both the bait (MD = 0.254, $p = 0.001$) and the visual (MD =

0.453, $p < 0.001$) thumbnails, while the bait thumbnail yielded a higher effect on dominance than the visual (MD = 0.199, $p = 0.011$) thumbnail. Dominance had been positively affected by activity on social media and sentiment towards COVID. This result indicates that participants who were more active on social media and had a more positive sentiment towards the pandemic reporting also experienced higher dominance while observing the social media posts.

Risk perception

Table 5 shows the perception of risk upon viewing the different types of visual stimuli has been successfully predicted by the type of stimulus, social media activity, hours spent on social media, and COVID sentiment. Participants estimated risk as higher upon viewing the visual thumbnail, in comparison to both the bait (MD = 0.099, $p < 0.001$) and the advice (MD = 0.122, $p < 0.001$) thumbnails, while there was no significant difference between the bait and the advice (MD = 0.022, $p = 1$) thumbnail. Activity on social media, hours spent on social media, and sentiment towards COVID had positive effects on risk perception. This result indicates that participants who were more active on social media, spent more time on social media, and had a more positive sentiment towards the pandemic reporting also experienced higher dominance while observing the social media posts.

Engagement

As Table 5 shows, engagement with posts, including the different types of visual stimuli, has been successfully predicted by social media activity and COVID sentiment. Both this activity on social media and sentiment towards COVID had positive effects on engagement. This result indicates that participants who were more active on social media and had a more positive sentiment towards the pandemic reporting also engaged more with COVID-related social media content.

Credibility

Table 5 shows the perception of the credibility of posts containing the different types of visual stimuli has been successfully predicted by the type of stimulus, activity on social media, and COVID sentiment. Participants estimated credibility as higher upon viewing the advice thumbnail, in comparison to both the bait (MD = 0.106, $p < 0.001$) and the visual (MD = 0.117, $p < 0.001$) thumbnails, while there was no significant difference between the bait and the visual (MD = 0.083, $p = 1$) thumbnail. Both activity on social media and sentiment towards COVID had positive effects on risk perception. This result indicates that participants who were more active on social media and had a more positive sentiment towards the pandemic reporting also perceived articles related to COVID as more credible.

Results summary

In summary, the participants seemed to have the most pleasant experience while observing the advisory visuals in the social media posts related to COVID-19. These posts yielded higher valence, arousal, dominance, credibility, and lower risk perception. Visual stimuli were not preferred, with the highest perception of risk and lower valence and dominance

than the other types of stimuli. The clickbait type of visual stimulus was most commonly in the middle between the other two, never being the one with the highest or the lowest scores on a measure. Furthermore, the visual type of stimuli was not relevant to the engagement at all. Lastly, activity on social media and sentiment towards COVID-19 successfully predicted all criteria, indicating that they are relevant for the emotions, risk perception, engagement, and credibility estimation of social media posts relating to COVID-19.

DISCUSSION

This study investigated the effect of different types of visual stimuli in a COVID-related social media post on users' reactions to that post. The study postulated six research questions to assess this effect, and the following part of the discussion will examine them one by one. Then, it will present an integration of all results along with the implications of the findings. This discussion will conclude with the study's limitations.

RQ1: What is the effect of the three types of visual stimuli on the valence of emotions that users feel after looking at a COVID-related social media post?

The answer to the first research question is that advisory visuals yielded the most positive emotional reactions, while the data visualization type yielded the most negative emotions. Clickbait-type visuals were between these two in this regard. This result may be a consequence of the data visualization stimuli leading the users to perceive the most risk, which scares them and initiates negative emotions. The advisory stimuli may simply be more pleasant to observe than the clickbait imagery, thus making them feel more positive. Aside from visual stimuli type, this study demonstrated that activity on social media and COVID-19 sentiment affect valence as well. The effect of the COVID-19 sentiment is likely a confirmation bias, with people who have more positive views on COVID-19 coverage and find it more important to have more positive reactions towards COVID-related news stories, similar to the findings of *Kim, Moravec & Dennis (2019)*. Furthermore, people who are more active on social media websites also experienced higher valence, which is likely due to a higher preference for viewing social media material, which leads them to feel higher valence in this study.

RQ2: What is the effect of the three types of visual stimuli on the arousal that users feel after looking at a COVID-related social media post?

The results of the study regarding arousal were similar to those regarding valence. Advisory imagery made the participants feel more aroused, and there were no differences in arousal between the other two types of imagery. Thus, advice on what to do during the pandemic seems to put the participants in a more awake, alert state, potentially by making them imagine themselves doing the recommended actions. Furthermore, like the previous topic, activity positively affected social media and COVID sentiment. The explanation for these factors may be the same as the previous question.

RQ3: What is the effect of the three types of visual stimuli on the dominance that users feel after looking at a COVID-related social media post?

Dominance exhibited an identical pattern of differences in the three types of imagery to valence: advisory imagery yielded the highest dominance, followed by bait, and data visualization in last place. This order is also logical, as the advisory imagery gives the

participant a sense of control or effect on their surroundings. On the contrary, data visualizations show large numbers of people affected by something, which may make the person feel tiny and out of control. As SAM is a non-verbal measurement, it may be that the advice, which puts them in the 'first plan' made them feel bigger, while the data put them in the background, making them feel smaller. Furthermore, feeling more positively toward COVID-19 reporting and spending more time on social media also had positive effects on dominance, likely due to feeling more in control in this specific situation of looking at a COVID-related social media post. People who look at social media posts more frequently will logically feel more dominant when looking at another one, because they are doing something they do frequently and confidently. The same is the case for those who feel positively towards COVID and are used to reading social media posts about it.

RQ4: What is the effect of the three types of visual stimuli on the users' perception of risk after looking at a COVID-related social media post?

Partially in line with the interpretation given for RQ1, the participants felt the most risk after looking at visualization-type imagery. As the introduction discussed, this type of imagery likely addresses risks explicitly by giving the users specific numbers (or visuals) which show how at-risk they, or some other people, are. Thus, it was natural that this type of imagery yielded the highest perception of risk. On the other hand, the remaining two types of thumbnails showed no differences in risk perception between them. This result is also in line with the findings on dominance, as the riskiest type of imagery is one in which the participants feel the least dominant. Previous research indicated that imagery in social media posts might be important for risk perception (*Bica et al., 2019*), which is in line with the findings of this study.

Those who are more active on social media and have a more positive sentiment towards COVID showed higher risk perception. This result indicates that they may trust social media more, as the credibility results have confirmed, which will be presented later. Since these participants trust the news on social media more, they take it more seriously and experience a higher perception of risk when interacting with it because it deals with the pandemic.

RQ5: What is the effect of the three types of visual stimuli on the users' engagement with a COVID-related social media post?

Engagement is the only study variable on which the types of imagery had no effect. The reason may be that the content of the social media post may be the deciding factor whether a person interacts with the post or not and not the visuals. Previous studies did show that altering the types of headlines changes engagement in various ways (*Janét, Richards & Landrum, 2022*), but instead of trying to manipulate the headlines, this study kept the headlines under control. Thus, the only recorded effects on engagement were those of social media activity, which is a measure of typical engagement, and of COVID sentiment, which may be another instance of confirmation bias, in which people who feel more positively towards something feel more inclined to interact with it. The lack of differences in engagement despite changes in arousal and other aspects of emotions may be due to the method by which the engagement was measured. The score on engagement measures the number of ways in which a participant is willing to engage with the content, rather than

their likelihood of engaging in general. As a result, it is possible that the measure was not sensitive enough to detect the subtle differences in users' likelihood of engaging based on the type of visual imagery and potentially provoked by their differences in emotional states.

RQ6: What is the effect of the three types of visual stimuli on the users' perception of the credibility of a COVID-related social media post?

The answer to the final research question is that advice seemed more credible than both clickbait and data visualization. This type of visualization may be the most straightforward, similar to how the traditional headlines were seen as the most credible in the study of *Janét, Richards & Landrum (2022)*. Their study also showed that agreeing with the topic of the post was instrumental in seeing it as more credible. This study arrived at the same findings due to the effect of COVID sentiment on credibility perception.

In sum, the results paint a clear picture: advisory thumbnails make the users the most positive, aroused, in control, have them not feel too much risk, and believe in the credibility of a social media COVID-related post the most. On the other hand, visual data imaging builds risk perception, negative emotions, a lack of control, and inferior credibility. Clickbait thumbnails fall mostly in between the other two. All these effects are persistent beyond the numerous control variables included in the models. This result is consistent with previous literature indicating that different aspects of social media posts have systematic effects on the audience (*Bica et al., 2019*; *Ghanem, Rosso & Rangel, 2020*; *Janét, Richards & Landrum, 2022*; *Kim, Moravec & Dennis, 2019*; *Kim & Dennis, 2019*).

Such outcomes may be due to the nature of these types of imagery—advisory imagery may provide users with a sense of control, or at least controllability. They are instructed to do certain activities that will benefit them, or at least keep them risk-free. That could explain why they feel more positively, more aroused, in control, and perceive risk as lower. They may also see the posts as more credible because they appear to be reasonable advice from relevant authorities. Data presentation imagery, on the other hand, has the opposite effect, which can be explained by its nature. When looking at large numbers, people may feel less like they have any effect. Therefore, they feel worse, less aroused, less in control, and perceive a high level of risk—they are afraid and helpless. Furthermore, the lower credibility may be due to common statistical manipulation in the media. These interpretations are slight exaggerations of reality, with only minor differences in scores, but they explain the likely mechanisms by which the differences arose.

Health officials and others who want to share information regarding COVID-19 on social media would be wise to utilize advisory thumbnails, to keep their participants feeling as positive as possible and enhance their credibility. On the other hand, if their wish is to scare the user and make them realize that they are in danger and that they do not have control over a situation, they should opt for various forms of data visualization. Future studies may want to integrate the investigations of headlines by *Janét, Richards & Landrum (2022)*. with examinations of imagery to determine the optimal conditions of information delivery in terms of how the users respond to it.

This study had two limitations. The first one was the usage of convenience sampling. The researcher's resources did not allow for random sampling, which would be more representative of the population of social media users. However, there is no reason to

believe that convenience sampling yielded a biased sample relevant to the study. Another limitation is that the study was not conducted in a naturalistic context—the participants did not observe the social media posts during their usual social media usage. Furthermore, the experiments in this study rely on questionnaires to assess the participants' perception of social media posts. As a result, the findings are less credible than those of objective data analyses or analyses of actual social media posts. However, it is impossible to systematically alter the thumbnail content while controlling for other factors using more naturalistic designs, which is why this study used the experimental approach. It maximized the ecological validity by replicating the exact look of Facebook social media posts.

CONCLUSIONS

This study investigated the effects of different types of thumbnails of COVID-related social media news posts on users' reactions. It determined that advisory imagery positively affected valence, arousal, dominance, credibility, and (lower) risk perception. Furthermore, imagery showing data visualizations yielded low valence, arousal, dominance, credibility, and high risk perception. Because of the low emotionality and high risk perception, data visualizations may be best suited for calmly warnings users of high risk situations. Clickbait-styled thumbnails which carry no information tended to fall between the other two types. The type of imagery did not affect the motivation to engage with a post. Aside from visual imagery, this study showed that most variables were affected by COVID sentiment and usual activity on social media. These results can help health officials and other users utilize imagery to deliver information about the COVID pandemic or other crises to cause specific emotions in the users.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The author received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Waleed M. Al-nuwaiser conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at figshare: Al-nuwaiser, Waleed (2022): Effect of visual imagery in COVID-19 social media posts on users' perception. figshare. Dataset. https://doi.org/10.6084/m9.figshare.20627823.v2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1153#supplemental-information.

REFERENCES

AlAzzam M, Abuhammad S, Abdalrahim A, Hamdan-Mansour AM. 2021. Predictors of depression and anxiety among senior high school students during COVID-19 pandemic: the context of home quarantine and online education. *The Journal of School Nursing* **37(4)**:241–248 DOI 10.1177/1059840520988548.

Allcott H, Gentzkow M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* **31(2)**:211–236 DOI 10.1257/jep.31.2.211.

Almandoz JP, Xie L, Schellinger JN, Mathew MS, Bismar N, Ofori A, Kukreja S, Schneider B, Vidot D, Messiah SE. 2021. Substance use, mental health and weight-related behaviours during the COVID-19 pandemic in people with obesity. *Clinical Obesity* **11(2)**:e12440 DOI 10.1111/cob.12440.

Alsubaie S, Temsah MH, Al-Eyadhy AA, Gossady I, Hasan GM, Al-Rabiaah A, Jamal AA, Alhaboob AA, Alsohime F, Somily AM. 2019. Middle East Respiratory Syndrome Coronavirus epidemic impact on healthcare workers' risk perceptions, work and personal lives. *The Journal of Infection in Developing Countries* **13(10)**:920–926 DOI 10.3855/jidc.11753.

Barthel M, Mitchell A, Holcomb J. 2016. Many Americans believe fake news is sowing confusion. Pew Research Center's Journalism Project. United States of America. *Available at https://policycommons.net/artifacts/618138/many-americans-believe-fake-news-is-sowing-confusion/1599054/* (accessed 29 October 2022).

BBC. 2020a. Coronavirus information: four posters. BBC. *Available at https://www.bbc.com/news/uk-51914645*.

BBC. 2020b. COVID: twelve charts on how COVID changed our lives. BBC. *Available at https://www.bbc.com/news/uk-55214991*.

Bica M, Demuth JL, Dykes JE, Palen L. 2019. Communicating hurricane risks: multi-method examination of risk imagery diffusion. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

Biro-Hannah E. 2021. Community adult mental health: mitigating the impact of Covid-19 through online art therapy. *International Journal of Art Therapy* **26(3)**:96–103 DOI 10.1080/17454832.2021.1894192.

Bradley MM, Lang PJ. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* **25(1)**:49–59 DOI 10.1016/0005-7916(94)90063-9.

Cao W, Fang Z, Hou G, Han M, Xu X, Dong J, Zheng J. 2020. The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Research* **287(10224)**:112934 DOI 10.1016/j.psychres.2020.112934.

Choi J, Lee SY, Ji SW. 2021. Engagement in emotional news on social media: intensity and type of emotions. *Journalism & Mass Communication Quarterly* **98(4)**:1017–1040 DOI 10.1177/1077699020959718.

Deutsch JI. 2017. Folklore, politics, and fake news in the reception of rogue one. *New Directions in Folklore* **15(1/2)**:109–120.

Dixon S. 2022. Mobile social media usage in the United States—Statistics & Facts. Statista*Available at https://www.statista.com/topics/3196/social-media-usage-in-the-united-states/* (accessed 10 August 2022).

Dormann C. 2001. Seducing consumers, evaluating emotions. In: *Joint Proceedings of IHM-HCI 2001*. Vol. 2. 10–14.

Dormann C. 2003. Affective experiences in the Home: measuring emotion. In: *HOIT*. Vol. 3.

Dryhurst S, Schneider CR, Kerr J, Freeman ALJ, Recchia G, van der Bles AM, Spiegelhalter D, van der Linden S. 2020. Risk perceptions of COVID-19 around the world. *Journal of Risk Research* **23(7–8)**:994–1006 DOI 10.1080/13669877.2020.1758193.

English R. 2017. Anti-vax truthiness: how social media echo chambers reinforce anti-vax beliefs in parents. In: *Social Media: Practices, Uses and Global Impact (Media and Communications-Technologies, Policies and Challenges)*. Hauppauge: Nova Science Publishers, 173–194.

Ghanem B, Rosso P, Rangel F. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)* **20(2)**:1–18 DOI 10.1145/3381750.

Grech V, Masukume G. 2017. Fake news of baby booms 9 months after major sporting events distorts the public's understanding of early human development science. *Early Human Development* **115**:16–17 DOI 10.1016/j.earlhumdev.2017.08.007.

Guess A, Nagler J, Tucker J. 2019. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5(1)**:eaau4586 DOI 10.1126/sciadv.aau4586.

Gui X, Kou Y, Pine KH, Chen Y. 2017. Managing uncertainty: using social media for risk assessment during a public health crisis. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4520–4533.

Isomursu M, Tähti M, Väinämö S, Kari Kuutti K. 2007. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies* **65(4)**:404–418 DOI 10.1016/j.ijhcs.2006.11.007.

Janét K, Richards O, Landrum AR. 2022. Headline format influences evaluation of, but not engagement with, environmental news. *Journalism Practice* **16(1)**:35–55 DOI 10.1080/17512786.2020.1805794.

Jo W, Lee J, Park J, Kim Y. 2020. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in South Korea: structural topic model and network analysis. *Journal of Medical Internet Research* **22(6)**:e19455 DOI 10.2196/19455.

Kim A, Dennis AR. 2019. Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly* **43(3)**:1025–1039 DOI 10.25300/MISQ/2019/15188.

Kim A, Moravec PL, Dennis AR. 2019. Combating fake news on social media with source ratings: the effects of user and expert reputation ratings. *Journal of Management Information Systems* **36(3)**:931–968 DOI 10.1080/07421222.2019.1628921.

Lang P. 1980. Behavioral treatment and bio-behavioral assessment: computer applications. In: *Technology in Mental Health Care Delivery Systems*. Norwood: Ablex, 119–137.

Liang L, Ren H, Cao R, Hu Y, Qin Z, Li C, Mei S. 2020. The effect of COVID-19 on youth mental health. *Psychiatric Quarterly* **91(3)**:841–852 DOI 10.1007/s11126-020-09744-3.

Mehrabian A. 1970. A semantic space for nonverbal behavior. *Journal of Consulting and Clinical Psychology* **35(2)**:248–257 DOI 10.1037/h0030083.

Mehrabian A, Russell JA. 1974. *An approach to environmental psychology*. Cambridge: MIT Press.

Osgood CE, Suci GJ, Tannenbaum PH. 1957. *The measurement of meaning. No. 47*. Champaign: University of Illinois Press.

Pietrabissa G, Volpi C, Bottacchi M, Bertuzzi V, Usubini AG, Löffler-Stastka H, Prevendar T. 2021. The impact of social isolation during the covid-19 pandemic on physical and mental health: the lived experience of adolescents with obesity and their caregivers. *International Journal of Environmental Research and Public Health* **18(6)**:3026 DOI 10.3390/ijerph18063026.

Pine KH, Lee M, Whitman SA, Chen Y, Henne K. 2021. Making sense of risk information amidst uncertainty: individuals' perceived risks associated with the COVID-19 pandemic. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

Qin H, Sanders C, Prasetyo Y, Syukron M, Prentice E. 2021. Exploring the dynamic relationships between risk perception and behavior in response to the Coronavirus Disease 2019 (COVID-19) outbreak. *Social Science & Medicine* **285**:114267 DOI 10.1016/j.socscimed.2021.114267.

Shane S. 2017. The fake Americans Russia created to influence the election. The New York Times. *Available at https://cs.brown.edu/people/jsavage/VotingProject/2017_09_07_NYT_TheFakeAmericansRussiaCreatedToInfluenceTheElection.pdf*.

Sutter JD. 2010. Texts, maps battle Haiti cholera outbreak. CNN. *Available at http://edition.cnn.com/2010/TECH/innovation/10/29/haiti.cholera.tech/index.html*.

Vahedian-Azimi A, Moayed MS, Rahimibashar F, Shojaei S, Ashtari S, Pourhoseingholi MA. 2020. Comparison of the severity of psychological distress among four groups of an Iranian population regarding COVID-19 pandemic. *BMC Psychiatry* **20(1)**:1–7 DOI 10.1186/s12888-020-02804-9.

Welbers K, Opgenhaffen M. 2019. Presenting news on social media: media logic in the communication style of newspapers on Facebook. *Digital Journalism* **7(1)**:45–62 DOI 10.1080/21670811.2018.1493939.

Westerman D, Spence PR, Van Der Heide B. 2014. Social media as information source: recency of updates and credibility of information. *Journal of Computer-Mediated Communication* **19(2)**:171–183 DOI 10.1111/jcc4.12041.

Wundt WM. 1913. *Grundriss der psychoxlogie*. Nile: A. Kröner.

PeerJ Computer Science

An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests

Hieu Nguyen and Swapna Gokhale

Computer Science and Engineering, University of Connecticut, Storrs, Connecticut, United States

ABSTRACT

Trust in the government is an important dimension of happiness according to the World Happiness Report (*Skelton, 2022*). Recently, social media platforms have been exploited to erode this trust by spreading hate-filled, violent, anti-government sentiment. This trend was amplified during the COVID-19 pandemic to protest the government-imposed, unpopular public health and safety measures to curb the spread of the coronavirus. Detection and demotion of anti-government rhetoric, especially during turbulent times such as the COVID-19 pandemic, can prevent the escalation of such sentiment into social unrest, physical violence, and turmoil. This article presents a classification framework to identify anti-government sentiment on Twitter during politically motivated, anti-lockdown protests that occurred in the capital of Michigan. From the tweets collected and labeled during the pair of protests, a rich set of features was computed from both structured and unstructured data. Employing feature engineering grounded in statistical, importance, and principal components analysis, subsets of these features are selected to train popular machine learning classifiers. The classifiers can efficiently detect tweets that promote an anti-government view with around 85% accuracy. With an F1-score of 0.82, the classifiers balance precision against recall, optimizing between false positives and false negatives. The classifiers thus demonstrate the feasibility of separating anti-government content from social media dialogue in a chaotic, emotionally charged real-life situation, and open opportunities for future research.

Subjects Computational Science, Data Mining and Machine Learning, Data Science
Keywords Social media, Anti-government, COVID-19, Lockdown protests, Machine Learning

Submitted 3 June 2022
Accepted 16 September 2022
Published 25 November 2022

Corresponding author
Swapna Gokhale,
swapna.gokhale@uconn.edu

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj-cs.1127**

Copyright
2022 Nguyen and Gokhale

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION & MOTIVATION

The World Happiness Report indicates that the extent to which people trust their governments and institutions plays an important role in their happiness and well-being (*Helliwell et al., 2021*). Communities which show high levels of trust are happier and more resilient in the face of a wide range of crises (*Skelton, 2022*). In the physical world, fringe groups that seek to question this trust by raising doubts in people's minds about the intent and motives of the government have always existed (*UN, 2012*). However, due to logistical reasons, these fringe groups in the offline era operated within a local scope, reaching only limited audiences. In the modern world, however, social media platforms have given easy, accessible, and approachable ways to these groups to spread their hateful, radical and

anti-government sentiment beyond the local boundaries of their operating areas to a far greater audience (*UN, 2012*).

The COVID-19 pandemic was one of the biggest health crises that we have seen in a century (*Skelton, 2022*). To control the spread of the virus and to keep people safe, state and local governments enacted many public health interventions such as masking, social distancing, and lockdowns. Of these, widespread lockdowns which ordered businesses and schools to shut down, and forced people to stay home were the most disruptive economically and socially. In the initial few weeks, many people viewed these orders as a necessary evil to protect our healthcare systems from being overwhelmed. As a result, they complied with these orders grudgingly, even though they viewed these draconian measures with skepticism.

A few weeks into the lockdown, however, anti-government sentiment fermented among far-right extremists. These extremist groups called for protests in many state capitals across the United States to oppose lockdown orders and to compel their governments to scale them back. These protests were considered ill-advised and unsafe by public health experts, and were covered extensively in national and international news.

Social media platforms offer a conduit for people to voice their opinions, thoughts and beliefs. Hence, these politically motivated and turbulent protests in the offline world also led to vigorous dialogue and exchange online as these platforms were the only channels of communication available to people to stave off their social isolation during the pandemic. Supporters of the protests often shared extreme and radical views, sowed distrust about governments' measures, contemplated ballot recalls to overthrow elected governments, and threatened violence against elected officials. Such anti-government rhetoric is often used for fundraising and recruitment purposes, and if left unchecked online, can lead to social unrest, violence and bloodshed. This was amply exemplified when subsequent protests involved guns, led to storming the Capitol building and eventually to a plot to kidnap the governor of Michigan (*Bogel-Burroughs, 2020*). However, if detected and mitigated earlier, such violent, out-of-control expression of anger and resentment may be prevented.

This article presents a classification framework to identify tweets which espouse extreme, anti-government views during politically charged protests in Lansing, Michigan. Tweets were collected during two separate anti-lockdown protests, and these tweets were then annotated as anti-government or non anti-government. A rich set of features was computed from both the structured and unstructured data collected along with these tweets. These features captured the content of the tweets, how Twitter users interacted with these tweets, and the properties of the tweets' authors. Three levels of feature engineering based on statistical significance, importance measures, and principal components was used to narrow down subsets of significant features that contribute meaningfully towards separating the tweets into anti-government and non anti-government groups. These subsets of features were used to train popular machine learning classifiers. The results showed that the classifiers could efficiently identify tweets that harbor anti-government sentiment with an accuracy of around 87%, with a training time of only a few seconds. The classifiers could also trade between precision and recall well with a F1-score of 0.82,

balancing between false positives and false negatives. These results are particularly noteworthy because unlike other studies where radical content is shared by a particular group which embraces a narrow philosophy (and hence most content either condemns or condones that philosophy), in this case anti-government tweets are shared within the broader context of COVID-19 measures, local and non-local politics, and the tactics employed during the protests. The results thus demonstrate that anti-government rhetoric can be separated from broad and general social media conversations, and opens opportunities for future research in this area.

The rest of the article is organized as follows: Section 2 compares related research. Section 3 summarizes the steps in data preparation. Section 4 presents computation of features. Section 5 provides an overview of classifiers and performance metrics. Section 6 discusses the results. Section 7 concludes the article and offers future directions.

RELATED RESEARCH

Presently, social media platforms have been exploited to share and spread radical and extreme ideas that sow suspicion and hatred in order to destabilize democratically-elected institutions and governments. Alongside, research to detect such content from social media feeds has also gained traction to stem this rising tide. In this section, we compare and contrast contemporary efforts that have appeared in the literature on the topic of identifying radical content from social media dialogue.

Miranda et al. (2020) describe a technique based on support vector machines to detect radicalism in the content shared on Twitter in Indonesia. *Qi et al. (2019)* analyze Twitter and Reddit data around Hong Kong protests to identify influencers. *Wolfowicz et al. (2021)* differentiate between Facebook profiles of violent and non-violent radicals. *Ahmad et al. (2019)* present a deep learning-based sentiment analysis technique to classify extremist and non-extremist tweets. At the level of user accounts, *Abd-Elaal, Badr & Mahdi (2020)* present a classification framework to detect ISIS and non-ISIS accounts on Twitter. Another study related to ISIS is by *Mussiraliyeva et al. (2020)*, where they seek to detect ISIS-related language in Kazak using ensemble learners. *Yasin et al. (2021)* use unsupervised k-means clustering to group tweets into extremist and non-extremist content. *Araque & Iglesias (2020)* detect radical content from ISIS accounts by comparing it against the content from news outlets such as the New York Times and CNN. *Wu & Gerber (2018)* explore the predictive power of social media data in determining the onset of civil uprisings during the Egyptian revolution, whereas a recent study summarizes the literature on this topic (*Grill, 2021*).

Most of the above studies have been conducted on radical extremism beyond the U.S. shores and in regions where this prevalence is believed to be high. In recent years, however, social media platforms have been exploited within the United States to erode trust in political institutions (*Nguyen & Othmeni, 2021*) by spreading extreme, off-mainstream content. A social media platform that has gained notoriety for sharing and propagating such content under the guise of free speech is Parler (*Aliapoulios et al., 2021*). Supporters of radical groups such as Proud Boys, Boogaloo Bois, and QAnon have also been particularly active in certain regions of the country and on platforms such as Twitter and Reddit, and

their activities offline and online have been studied (*DeCook, 2018*; *Klein, 2019*; *Reid, Valasik & Bagavati, 2020*; *Fahim & Gokhale, 2021*). These works focus on radical and extremist content shared by groups that espouse a particular cause or a philosophy, and hence, most of the supporting chatter either promotes or criticizes that philosophy. For example, users of Parler are concerned about free speech, and Proud Boys is a far-right, neo-fascist, exclusively male organization.

In our work, however, anti-government discourse is embedded in a broader context, ranging from protesting COVID-19 public health measures to campaigning for (or against) local and non-local politicians and their governing philosophies, to criticizing (condoning) the tactics employed during the protests. We build a framework that can mine social media feeds to provide unique and early insights into people's radical opinions and thoughts, regardless of such context. This framework can help prevent violence and social unrest, especially in turbulent and chaotic circumstances such as those brought about by the COVID-19 pandemic, which was amply exploited by extremists to spread mistrust and hatred about the government and its policies (*Clarke, 2022*).

DATA PREPARATION

This section discusses three steps in the preparation of data: collection, annotation, and pre-processing.

Data collection

During the months of April and May 2020, anti-lockdown protests were organized in many states, including North Carolina, Michigan, Pennsylvania, Virginia, and California (*Andone, 2020*). Of these, the protests that occurred in Michigan gained ill reputation for many reasons. First, the crowds of protesters in other states were in the hundreds while the Michigan protests were the largest, attracting thousands of protesters. Second, while protesters in other states simply gathered on the streets, those in Michigan engaged in many violent and questionable tactics that threatened the health and well-being of many. Third, the protesters wanted to draw attention to the conflicting political situation in Michigan and its status as a battleground state in the 2020 presidential election. This conflict arose because Michigan was led by a Democratic governor but voted for President Trump in the prior (2016) presidential election. Thus, the protesters included members of Women for Trump, mainstream Republicans, anti-vaxx and gun rights advocates, Proud Boys, and Boogaloo Bois (*Ecarma, 2020*; *Wilson, 2020*). These factors drew media attention, followed by considerable volumes of conversations on social media platforms. Therefore, although protests were conducted in many states, we chose Michigan as a prominent example of anti-lockdown protests in the U.S. The specific circumstances surrounding the two protests, which were considered in building a coding guide for the annotation of tweets, were as follows:

- **Operation Gridlock:** This was the name given to the first protest. It was organized by a Facebook group with the same name, created by the Michigan Freedom Fund and Michigan Conservative Coalition. Close to 3,000 people showed up, the protest lasted

8 h, and the protesters blocked ambulances from reaching the only Level I trauma center at Sparrow Hospital. Most stayed in their cars, jammed the streets around the capitol building, and caused delays during a shift change at the hospital. About 150 protesters spilled on the lawn of the Capitol, flouting social distancing and masking guidelines. Protesters carried confederate, Nazi, and American flags (*Berg & Egan, 2020*).

- **Michigan Protest:** During the second Michigan protest, hundreds of protesters carried firearms, dressed in camouflage and military garb, gathered at the Capitol, and many managed to enter the building. Thus, the second protest took a more violent tone. It was organized by the conservative group American Patriot Council. Confederate flags, swastikas, and nooses were present at this protest too (*Mauger, 2020*).

We collected tweets a few days following the two Michigan protests on April 15, 2020, and April 30, 2020. Corresponding to the respective trending hashtags, we used *#operationgridlock* for the first sample and *#michiganprotests* for the second sample. The tweets were collected using the rtweet API (*Kearney et al., 2020*). Each time, the sample resulted in about 4,000 tweets.

Data annotation

The objective of our research is to identify anti-government, deviant content from the rest of the dialogue because such content seeks to undermine the faith and trust in the government and its policies. This loss of trust may make the government's job of protecting the people considerably harder. Therefore, to study our research question, we chose to label the tweets into two groups, one group consisted of anti-government tweets, and the second group included tweets that are not against the government or non anti-government. We note that the second group may contain a mix of pro-government and neutral tweets, and if our research question were stance detection (*Cotfas et al., 2021*), we would further split the second group into these two categories. However, because the scope of our research question is limited to detecting tweets that sow resentment and suspicion against the government, we chose to combine the tweets that voiced support for the government along with the neutral tweets. Two additional reasons also motivated us to retain the pro-government and neutral tweets into a single class. First, generally, pro-government content may be considered suspicious and propagandist in autocratic or non-democratic regimes (*Stukal et al., 2022*; *Caldarelli et al., 2020*). However, governments in the U.S. at all levels (local, state, and federal) are elected democratically through free and fair elections. Thus, in this dialogue, pro-government content affirmed support for the public health restrictions that were implemented and was not viewed as propaganda. Second, our human coders found separating between these two types of tweets confusing, yielding a lower agreement between them. Therefore, we sought to annotate each tweet as either anti-government ('A') or non anti-government ('N').

To facilitate this annotation, we built a coding guide that the manual annotators could consult. This coding guide consisted of themes and the representative examples of both anti-government and non anti-government tweets that fitted each theme. Through an

extensive review of the news stories and opinion pieces, it was observed that most of the tweets fell along the following themes:

- **Tactics and Circumstances:** These tweets referred to the tactics employed by the protesters, and the other circumstances surrounding the protests. Although these tactics were disruptive and even violent; naturally, anti-government tweets praised them for the inconvenience they caused and the threatening/intimidation situations they produced. On the other hand, non anti-government tweets condemned them for the same reasons.
- **Local Politics:** These tweets mentioned local political figures in Michigan, with Governor Whitmer appearing predominantly. DeVos was another highly visible Republican family in Michigan with a significant presence and was believed to have sponsored the protests (*Hernandez, 2020*). Anti-government tweets denigrated the governor as a dictator and a Nazi, whereas non anti-government tweets stood with her in solidarity.
- **Non-local Politics:** These tweets cast the protests in Michigan as a part of the broader landscape and encouraged people in other states and nationally to engage in similar resistance and rallies to ease COVID-19 restrictions. Nationally visible Republican and Democratic leaders and governors of other states were mentioned in these tweets.
- **COVID-19:** These tweets explicitly referred to COVID-19. Anti-government tweets questioned the motive behind the public health measures and expressed skepticism about the seriousness of the virus. Non anti-government tweets mostly voiced concern about how these protests, which also came with rebelling against the public health guidelines such as social distancing and masks, would affect the trajectory of the number of cases.
- **Political Ideology:** These tweets were ideologically inspired; anti-government tweets praised the protesters as patriots and defenders of individual liberties and freedoms, while non anti-government tweets were critical of the protesters as white supremacists and racists.

Tables 1 and 2 show representative examples of anti-government and non anti-government tweets for each theme for Operational Gridlock and Michigan Protest data sets respectively. This coding guide was given to two annotators, who labeled each tweet as either anti-government ('A') or non anti-government ('N'). We eliminated duplicates before labeling. Both annotators had to agree upon the label for a tweet to be included in the final *corpus*. The disagreement between the two annotators eliminated approximately 450 tweets from each data set. The coders only coded each tweet as either anti-government or non anti-government; they did not identify the specific theme associated with the tweet. Therefore, although tweets from all the five themes were included in the analysis, it is not feasible to provide the split of the tweets into these five themes. The collective summary of the tweets from all five themes and their distribution between anti-government and non anti-government groups in the individual and the combined data set is summarized in Table 3.

Table 1 Themes & example tweets—Operation Gridlock.

Theme I: Tactics & circumstances	
A	This doesn't even begin to show the number of people in Lansing. Block and blocks of people siting in there cars. It didn't look like so many by the capital bc some streets were clo we d off. Vast majority stayed in their cars.
N	A friend took this from a hospital in Lansing, Michigan. Apparently the protesters blocked an ambulance from getting to the hospital. https://t.co/RxpA9S4TvL
Theme II: Non-local politics	
A	@GovInslee Jaydid you pay attention to Lansing tonight?
N	Yet these #Trump supporters call themselves #ProLife—blocking Hospital Workers from getting to work. #Lansing https://t.co/Tl42VMk22E
Theme III: Local politics	
A	I'm thinking Lansing MI doesn't like their gov and her non-essential bullshit???
N	@GovWhitmer I stand with Governor Whitmer. It's nice to have a Governor actually care about the people of her state!!! I'm embarrassed by the selfish assholes protesting in Lansing today.
Theme IV: COVID-19	
A	TRUTH IN NUMBERS. FLU BEING REPLACED FOR COVID $$$ LIVE SHUTDOWN PROTESTS IN LANSING MICHIGAN #COVID1984 #EndTheLockdown #F…
N	Presumably Lansing will now be Michigan's next virus hot spot.
Theme V: Political ideology	
A	#MichiganProtest is not about going out to eat or getting a haircut—its about govt restricting our #NaturalRights to #Freedom #Liberty. @NatlGovsAssoc #gretchenwhitmerisa #Fascist #oppressor @GovWhitmer #ProtestLockdown #ProtestTyranny #WeveHadEnough
N	If this virus has taught me anything its that Americans literally have zero fucking clue what the constitution actually says. #COVID #MichiganProtest

FEATURE COMPUTATION

The Twitter API returns both structured and unstructured data that represents the properties of the tweets and their authors in addition to the text, which is, of course, the core content of the tweet. Figure 1 shows the high-level processing pipeline of the steps involved in taking this raw data and converting it to features. Broadly speaking, we have three types of data, the text of the tweets, the parameters representing how Twitter users interacted with these tweets, and the inherent characteristics and activity level of the authors. In the next subsections, we elaborate on how each of these different types of data were mapped to features using Fig. 1 as the guide. For the numerical features, statistical significance between the two groups is assessed using the two-sample t-test (*Zafarani, Abbasi & Liu, 2014*).

Text features

The text of the tweets is the most important piece of the data. Therefore, it is no surprise that the text's content, its syntactical presentation, and the hidden underlying emotion all contribute to determining whether a tweet is anti-government or otherwise.

The content of the tweet conveys the actual message of the author. This message is revealed only after the noise is separated from the tweet, which was achieved through the following pre-processing steps. First, we removed the uninformative symbols such as mentions, hashtags, and hyperlinks. After this step, words that comprised the hashtag and mentions were regarded as a part of the tweet's text. Then, we eliminated stop words that

Table 2 Themes & example tweets—Michigan protest.

Theme I: Tactics & circumstances	
A	#Patriots, please use some restraint.. Whitmer is trying to goad Patriots into violence so she can justify #gun grab. #MichiganProtest #2A #OathKeepers @TheJusticeDept
N	My better judgement tells me that if you show up at a State Capital building decked out in riot gear and an AR 15, you would be immediately arrested, locked up and charged with assault? What happened to those days?? #MichiganProtest https://t.co/slDvRVNYod
Theme II: Non-local politics	
A	Leave it a vile #Democrat to equate protesting to racism. #LiberalismIsAMentalDisease #OPENAMERICANOW #MichiganProtest https://t.co/GSPqpIsQ7n
N	Why does @realDonaldTrump have a habit of calling people with confederate flags, swastikas, and nooses "Very fine people" or "very good people." Hmm. #Charlottesville #MichiganProtest.
Theme III: Local politics	
A	Fuck off @GovWhitmer, you're trash and so are your politics. #MichiganProtest #BidenTheRapist #MeTooUnlessItsBiden #MeToo https://t.co/EL7weMR9Fb
N	Flint hasn't had drinking water for a decade but don't make #Michiganers stay home for a month! #MichiganProtest #worthless #loserswithguns
Theme IV: COVID-19	
A	@GovWhitmer awful response to COVID will dwarf her confusion to her self-induced Michigan SHUTDOWN RECESSION plan—a massive train wreck. #COVID19 #MichiganProtest
N	@VP was reportedly told by hospital officials that a mask was a requirement and he still refused to wear one, while visiting #COVID_19 patients. To feed into the redoric of violant #DomesticTerrorists that participated in last nights #MichiganProtest for
Theme V: Political ideology	
A	All I can say is God Bless those patriots who grabbed their AK-47, filled up their Dodge Ram with their stimulus check and headed to Lansing to ignore the CDCs social distancing guidelines! These are the REAL heroes!
N	But we have dummies in Lansing marching for white supremacy…smh

Table 3 Summary of data sets.

Protest	Total	Anti-government	Not Anti-government
Operation Gridlock	3,570	931 (26.08%)	2,639 (73.90%)
Michigan protest	3,596	956 (26.59%)	2,990 (83.15%)
Total	7,166	1,887 (26.33%)	5,629 (78.55%)

commonly occur in the English language. These are the words that appear in NLTK's English stop word dictionary. We also added words such as "u" and "ur"; these words are commonly found in tweets but are not in the English dictionary. Added to this list of stop words were context-specific words such as "lansing", "michigan", "people" and "today", which occurred with similar frequency in both groups and hence, were likely to not contribute to the classification. Finally, the words were stemmed down to their stems using the PorterStemmer in the NLTK library (*Willett, 2006*). Stemming involved truncating the words down to their roots, for example, "writing", "writes", "written" got truncated to "write". The text remaining after pre-processing was then split into a list of words.

Figure 2 shows the word cloud visualizations built from the topmost unique words (remaining after pre-processing) in anti-government and non anti-government tweets. Anti-government tweets promote the ideological perspective that lockdowns are an

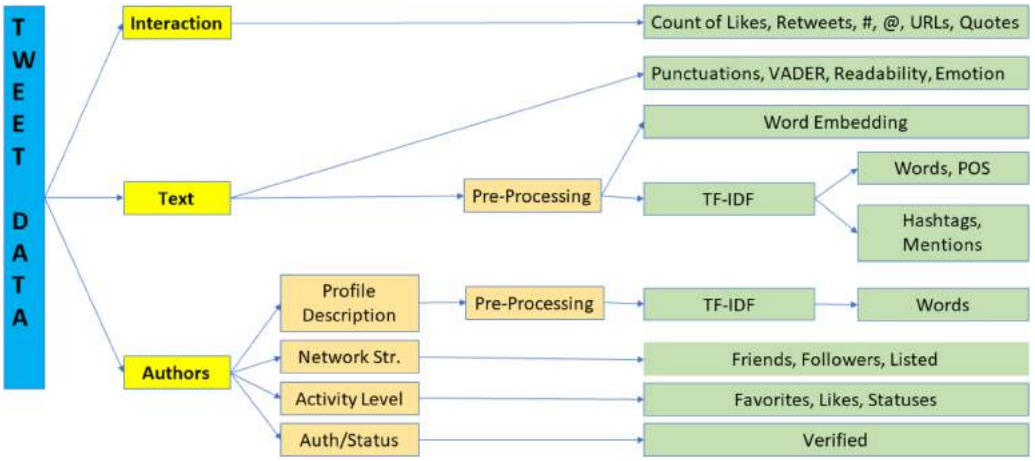


Figure 1 **Feature computation processing pipeline.** Full-size DOI: 10.7717/peerj-cs.1127/fig-1

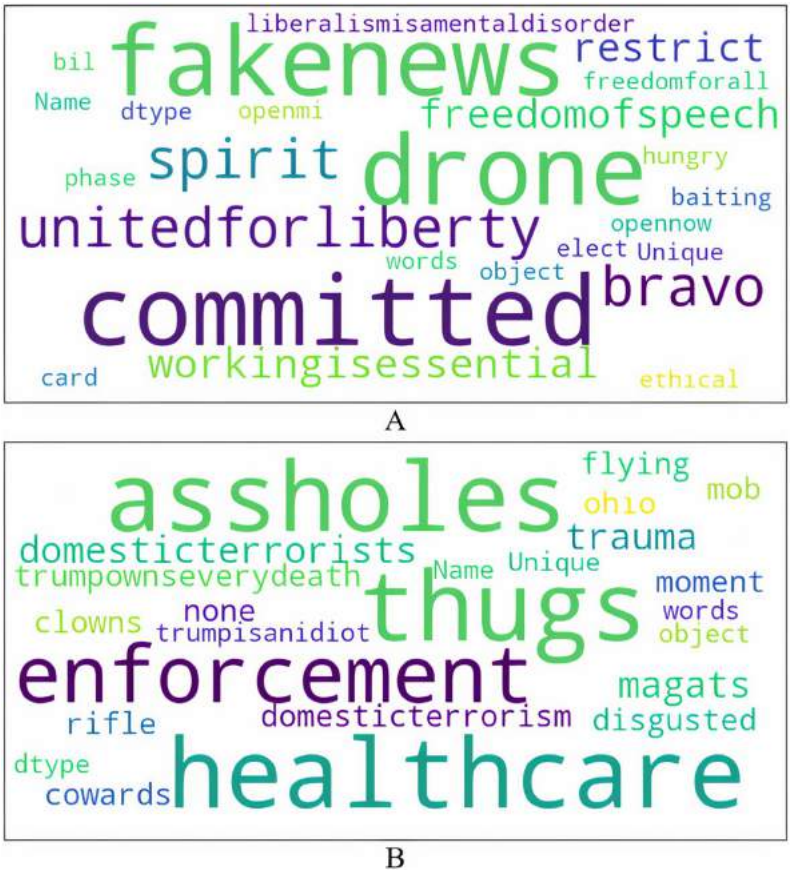


Figure 2 **Unique words in tweets.** Full-size DOI: 10.7717/peerj-cs.1127/fig-2

infringement on individuals' liberty and freedom. They mock liberalism, call for an end to the lockdowns by claiming that working is an important economic activity (all non-essential businesses were ordered to be shut down), and embrace skepticism about the virus. On the other hand, tweets that are non anti-government insult the protesters by

referring to them as "assholes", "thugs", "clowns", "magats", "cowards", "mob", and "domestic terrorists". Some tweets also associate the protesters with President Trump and his political views as suggested by these examples. In the tweet *I live in Lansing…why I've seen is armed protesters…many of whom are KKK, Michigan militia and assorted assholes*, the term "assholes" is used in the context of Michigan militia that aligned with the president. Another tweet *These #maga thugs are blocking the entrance to Sparrow Hospital in Lansing, MI to protest the stay at home order, calling it Operation Gridlock* refers to the protesters as MAGA thugs. MAGA (Make America Great Again) was President Trump's election slogan. The president is also blamed for COVID deaths as seen by "trumpownseverydeath". There are many tweets that lament over the large number of deaths due to COVID, further alluding to the overall apathy and indifference of the president and his supporters to this issue; for example *Only 153 deaths today? Sounds like a great day for a Trump rally in Lansing. Sorry idiots Trump isn't there today. #OperationGridlock*. Disgust and disbelief in these ill-informed protests asserting freedom is also expressed.

We used *n*-grams/TF-IDF (Term Frequency Inverse Document Frequency) and word embeddings to map these unique words to features. In the *n*-grams method, a text sample is represented by the most frequent instances of every unique *n* continuous words as a dimension. We calculated bigrams from the pre-processed text, and the weight for each bigram was its TF-IDF score (*Zafarani, Abbasi & Liu, 2014*). Words that make up hashtags and mentions were also included in the computation of bigrams, and hence, TF-IDF scores. TF-IDF scores assign a higher weight to those bigrams that are the most important differentiators between anti-government and non anti-government tweets. We calculated TF-IDF vectors from our *corpus* for top the 2,000 relevant bigrams.

TF-IDF provides a weighted score based on statistical importance. However, it does not preserve the contextual relationship among the words. To represent the context between the words, we computed Word2Vec embeddings based on neural networks that map semantically related words to low-dimensional, non-sparse vectors (*Mikolov et al., 2013*). Using Gensim library (*Rehurek & Sojka, 2010*), our model mapped each word to a 10-dimensional vector, with a minimum count of 1, and the number of partitions during testing set to 8. Because our data size was small, we used the skip-gram model (*Nicholson, 2019*).

The syntax and other arrangements of the words in the tweets capture how the content is conveyed. Authors may use various punctuations such as question marks, exclamations, quotes, etc., and other markers such as emoticons and upper case letters to emphasize how strongly they feel about their content (*Araque & Iglesias, 2020*). They may also use higher proportions of certain parts-of-speech such as adverbs and verbs to express their passion (*Xu, 2014*). In face-to-face communication, facial expressions and body language usually provide additional clues about the underlying emotion and passion of the speakers and their intensity. These clues, however, are not present in written communications, including social media texts. Therefore, in written texts, these syntactical patterns and organizations of words can substitute for facial gestures and non-verbal clues. Moreover, prior research indicates that these non-textual parameters may differ between polarising and regular

tweets (*Araque & Iglesias, 2020*). To represent these features, we included counts of question marks, exclamation marks, periods, quotation marks, links, and capital words. We also computed TF-IDF scores for different parts of-speech (*Xu, 2014*), using the NLTK library (*Loper & Bird, 2002*).

We computed two readability scores, representing the ease with which readers can understand the tweets. Readability is determined by how complex the vocabulary is, its syntax, and how the content is organized into sentences and paragraphs (*Benoit et al., 2018*). These are the Flesch Reading Ease and Flesch-Kincaid Grade Level indices (*Talburt, 1985*), and their respective expressions are:

$$\text{Flesch Reading Ease Index} = 206.835 - 1.1015\left(\frac{\text{W}}{\text{S}}\right) - 84.6\left(\frac{\text{L}}{\text{W}}\right) \tag{1}$$

$$\text{Flesch Grade Level Index} = 0.39\left(\frac{\text{W}}{\text{S}}\right) + 118.0\left(\frac{\text{L}}{\text{W}}\right) - 15.59 \tag{2}$$

In these indices, S represents the total number of sentences, W represents the total number of words, and L represents the total number of syllables in the text for which the indices are to be computed. W captures the sentence length, and L captures the word length. The philosophy behind using sentence length and word length is that longer sentences and words are more difficult to read and understand than shorter sentences and words.

Each index weighs these two factors differently, because of the different outputs they produce. The Flesch Reading Ease index produces a score between 0 and 100, which is then interpreted in 10-point intervals to determine readability. The higher the score, the easier it is to read the text, with scores between 100.0 and 90.0 suitable for a 5th grader, whereas scores between 10.0 and 0.0 suitable for a professional. On the other hand, the Flesch Grade Level index directly presents the score in terms of suitability for U.S. grade level (*Talburt, 1985*). Table 4 lists the values of these readability indices for anti-government and non anti-government tweets. As indicated by the p-values, there was a statistically significant difference in the Flesch Reading Ease index between the two groups, but the difference in the Flesch-Kinacid Grade Level index was statistically insignificant.

Tweets with disruptive information generally exhibit less emotion and sentiment and tend to be overall negative. By contrast, regular tweets that voice support for the democratic institutions may have a positive outlook and sentiment (*Araque & Iglesias, 2020*). These differences were quantified using the VADER sentiment scores computed from the text of the tweets (*Hutto & Gilbert, 2014*). The positive, negative, neutral, and compound scores for the tweets from the two classes are listed in Table 4. The difference in all the scores was statistically significant between the two groups. We also used scores for six emotions; namely, sadness, joy, love, fear, anger, and surprise computed using a pre-trained DistilBERT model. This model is a fast, cheap light transformer model based on the BERT architecture. The model is trained on the Twitter sentiment analysis data set and shows an accuracy of 93.8% and F1-score of 93.79% (*Saravia et al., 2018*). Table 4 shows that scores for the emotions of sadness, joy, and anger were statistically significant, whereas

Table 4 Readability, sentiment & emotion features.

Readability			
Parameter	A	N	*p*-value
Flesh reading ease	68.625	72.916	0.0000
Flesch-kincaid grade level	9.283	9.100	0.1845
VADER sentiment			
Parameter	A	N	*p*-value
Vader negative	0.157	0.212	0.0000
Vader positive	0.088	0.071	0.0000
Vader neutral	0.754	0.717	0.0000
Vader compound	−0.156	−0.353	0.0000
Emotions			
Parameter	A	N	*p*-value
Sadness	0.046	0.062	0.0008
Joy	0.176	0.134	0.0000
Love	0.003	0.003	0.9140
Anger	0.737	0.764	0.0046
Fear	0.035	0.034	0.8253
Surprise	0.002	0.002	0.9252

scores for the emotions of love, fear and surprise were not significant between the two groups.

As shown in Fig. 1; sentiment, readability, and emotion scores were computed from the raw text prior to pre-processing, which is why they are all listed together in Table 4.

Interaction features

When users tweet their thoughts, they expect other users to engage with and react to their tweets. Twitter users interact with the tweets in two public ways, liking (favoriting) and retweeting. Both likes and retweets may be viewed as forms of endorsement. Likes may be considered as a more tacit, passive method where the message contained in a tweet resonates with the user. Tweets liked by users are visible to their friends. On the other hand, retweeting is an active approach where users re-broadcast the tweets that their friends share to their entire network. Retweeting could be used to show the intention of listening and agreeing with the tweet owner's point of view (*Tweettabs, 2022*; *Hajibagheri & Sukhthankar, 2014*). Both these forms of interactions boost the diffusion and spread of the tweets, and therefore we use these parameters to measure the degree of interactions with a tweet. The table also compares these parameters for quoted tweets. Table 5 shows that average number of likes and retweets is significantly higher for anti-government compared to non anti-government tweets. The same trend holds for quoted tweets, and these differences are significant at the 5% level.

Users may also deliberately engineer their tweets to improve interaction (*Oehmichen et al., 2019*). We discuss the meaning of these actions and whether there exist any

Table 5 Interaction features.

Parameter	A	N	*p*-value
#likes	81.37	12.98	0.0004
#retweets	23.82	5.69	0.0017
#likes (Q)	21,918.82	7,729	0.0000
#retweets (Q)	6,385.70	2,900.70	0.0000
#hashtags	1.13	1.24	0.0210
#mentions	0.49	0.78	0.0000
#quotes	0.13	0.19	0.0158

quantifiable differences in these acts between the two groups. First, users may annotate tweets with one or more hashtags; adding a hashtag to a tweet allows other users to find their tweets and to interact with them. Adding a hashtag also builds a community of users discussing the same topics. Twitter also uses hashtags to calculate trending topics of the day, which further encourages users to post and join these communities (*Hajibagheri & Sukhthankar, 2014*). In the table, we show the average number of hashtags per tweet for both groups. Users may also mention other users; this can be viewed as a social activity, where one user is taking account of other user(s), and is oriented towards the course. Mentions can also be used to improve visibility. Finally, the tweets may also quote other tweets with additional comments of their own. These comments may either support the content of the original tweet or refute it. Either way, it can enhance the visibility of their own tweet, especially if the quoted tweet is from a prominent or a verified account (*Park, Compton & Lu, 2015*). Finally, authors may also craft URLs into their tweets to support their point of view with scientific or literary evidence or news articles. Although the magnitude of the difference in the numbers of hashtags, mentions, and quotes did not appear too large between the two groups, the difference was still significant at the level of 5%.

Authors' features

Data regarding the authors of the tweets is of two types; their profile descriptions and how they behave over the platform in terms of sharing their own thoughts as well as responding to others' content. We extracted features from both these types of data because prior research shows that both these factors, namely, whom the authors claim to be and how they act, will influence how far their tweets will spread (*Mann, Gaines & Gokhale, 2022*; *Oehmichen et al., 2019*).

The profiles of the authors were pre-processed through the same pipeline used for the text of the tweets. Word cloud visualizations were built from the resulting text, and TF-IDF features representing these profiles were extracted. These word clouds for authors who tweet anti-government and non anti-government content are shown in Fig. 3. The few words that stand out from the profiles of authors who shared anti-government tweets include *conserv*, *american presid*, *god*, *maga* and *family*. These words indicate that authors with conservative ideology, harboring support for the constitution, faith and family shared



Figure 3 **Author profiles of tweets.** Full-size DOI: 10.7717/peerj-cs.1127/fig-3

anti-government tweets. They may also be ardent supporters of President Trump and his populist campaign theme MAGA. On the other hand, prominent words in the word cloud of non anti-government tweets' authors are *human*, *social*, *democrat*, *artist*, *teacher*, *author*, *advocate*. These words point to those whose philosophy is liberal-leaning, may be employed in service-oriented professions, and are advocates of social justice and causes.

Next, we divided the parameters reflecting authors' behavior into those that represent the strength of their network, their authenticity and status, and their level of activity. Wherever available, parameters in these three groups were also compared for authors of quoted tweets. These three groups of parameters are discussed below (in this discussion original tweet refers to the tweet sampled by the API that appears in the *corpus*), and their average values for both anti-government and non anti-government tweets are in Table 6:

- **Network strength:** The strength of the authors' network can be assessed by the numbers of friends and followers. Generally, tweets of those authors who have larger networks of friends and followers can expect greater interaction and visibility. The numbers of friends and followers are compared both for authors of original and quoted tweets. The listed count indicates the number of other users who have added an author to their list and can be an indicator of popularity (*Tweettabs, 2022*). It is thus reasonable to believe

Table 6 Authors' features.

Network strength			
Parameter	A	N	*p*-value
#Friends	2,373.93	3,338.45	0.0014
#Followers	6,854.95	6,115.54	0.7929
#Friends (Q)	19,146.97	23,073.83	0.6044
#Followers (Q)	413,655.74	1,766,298.38	0.0520
#Lists	58.96	46.37	0.2013
Activity level			
Parameter	A	N	*p*-value
#statuses	26,542.22	31,987.65	0.0480
#favorites	27,609.38	25,473.09	0.2678
#statuses (Q)	42,560.14	51,456.97	0.2129
Authenticity/Trust			
Parameter	A	N	*p*-value
% Is verified	0.0135	0.0257	0.005

that tweets of authors with greater listed counts will be more popular and will receive more likes and retweets. The table indicates that only two parameters, namely, the number of friends of the authors of original tweets and the number of followers of the authors of quoted tweets, are significant. The difference in the other parameters is insignificant between the two groups.

- **Activity level:** One of the main indicators of the degree to which the authors are active on the platform is the number of status updates they have shared through the entire period that their accounts have been active. Status updates were compared for the authors of both the original and quoted tweets. Authors of non anti-government tweets have posted a significantly greater number of status updates compared to the authors of anti-government tweets. However, this difference is insignificant for authors of quoted tweets from both groups. Other secondary indicators include the number of times they have liked tweets from their friends and followers. Authors who prolifically react to tweets that appear on their feeds are likely to invite similar altruistically reciprocal relationships from their friends and followers (*Oehmichen et al., 2019*). Thus, the number of likes a tweet receives may have a high positive correlation with the number of tweets the author may have liked. However, there is no significant difference in the number of likes (listed as the number of favorites in Table 6 according to the nomenclature used by the Twitter API) by the authors of tweets from both groups.
- **Authenticity/Trust:** Tweets from high-profile, celebrated authors may attract a lot more attention, probably because the general public implicitly believes that the content shared from their accounts is more trustworthy and authentic. Moreover, these authors tend to have much larger networks of followers than those who are not celebrities. These popular authors who enjoy celebrity status tend to have accounts verified by Twitter; and hence, whether a tweet is shared from a verified account can be a factor in

influencing its spread. Thus, the table also compares the percentage of tweets shared from verified accounts for both classes. In the absolute sense, the percentage shared from verified accounts is trivial for both anti-government and non anti-government tweets. However, the difference between the percentage is statistically significant, as indicated by the *p*-value.

CLASSIFIERS AND PERFORMANCE

We chose the following popular models for classification. Multiple models are used since each model uses a different philosophy to arrive at a decision, and it is impossible to tell *a priori* which model will offer the best performance for a specific data set. The collection of models included an ensemble learner (Random Forest), a simple, basic learner (Logistic Regression), a simple, sophisticated learner (Support Vector Machine), a neural network (Multi-Layer Perceptron), and a pre-trained transformer model (DistillBERT). Model implementations in the Scikit package were used (*Pedregosa et al., 2011*), and their hyperparameters are listed below:

- **Random Forests (RF):** Random Forests is an ensemble learning method, where the underlying weak learner is a Decision Tree (*Liaw & Wiener, 2002*). It uses bagging to reduce variance by generating a number of decision trees with different training sets and parameters. The parameters of the model are as follows. Each forest consisted of 100 trees, the maximum number of features used to grow each tree in the forest is set to the square root of the total number of features (approximately 25–30 when all the features are employed), and each decision tree is not pruned.
- **Support Vector Machines (SVM):** Support Vector Machines (SVM) is a powerful classification technique that estimates the boundary (called hyper-plane) with the maximum margin (*Suykens, Lukas & Vandewalle, 2000*). We used SVMs with both linear and RBF kernels and L2 regularization. The regularization parameter C was set to 1, and kernel coefficient γ was set to scale.
- **Logistic Regression (LR):** One of the basic and popular algorithms to solve classification problems, this is named as such because of the Logit function that forms its basis. The parameters are penalty–L1, tolerance for the stopping criteria–0.0001, the inverse of the regularization strength C–1.00 and the maximum number of iterations–100 (*Pedregosa et al., 2011*).
- **Multi-Layer Perceptron (MLP):** Multi-Layer Perceptron is a feed-forward Artificial Neural Network (ANN) that consisted of input, hidden, and output layers (*Delashmit & Manry, 2005*), set to 10, 8, 5 and 2 respectively. We used rectifier linear unit (ReLu) instead of the sigmoid activation function to handle the problem that the derivative of the activation function rapidly approaches zero. This problem with the derivative is common in deep neural networks.
- **DistilBERT (D-BERT):** BERT (Bidirectional Encoder Representation from Transformers) is a deep learning model in which all outputs are connected with each input, and the weightings between them are dynamically calculated in the attention

layers (*Devlin et al., 2018*). This characteristic allows the model to understand the context of the words based on their surrounding words as compared to directional NLP models. We employed DistilBERT, a compact version of BERT where the model has 40% fewer parameters than BERT while preserving over 95% of BERT's performance (*Sanh et al., 2019*). The parameters of the DistilBERT model include: vocabulary size (30,522), max position embeddings (6), number of layers (6), number of heads (12), dimensions (768), number of hidden dimensions (3,072), dropout (0.1), attention drop out (0.1) and activation function (gelu) (*Sanh et al., 2019*).

Our main objective was to identify anti-government tweets; hence, to define the performance metrics, we designated the anti-government and non anti-government classes as positive and negative, respectively. Tweets could thus be classified into four groups–true positive (TP) (anti-government labeled as anti-government), true negative (TN) (non anti-government labeled as non anti-government), false positive (FP) (not anti-government labeled as anti-government), and false negative (FN) (anti-government labeled as not anti-government). These four groups led to the following metrics to compare classifier performance:

- **Accuracy (A):** Accuracy was defined as the percentage of tweets that are labeled correctly.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

- **Precision (P):** Precision measured the percentage of tweets that were actually anti-government out of all the tweets that were predicted as anti-government.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

- **Recall (R):** Recall measured how many of the anti-government tweets were actually labeled as anti-government.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

- **F1-score (F1):** F1-score balanced between Precision and Recall.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Precision is the percentage relevant from the set detected and recall is the percent relevant from within the global population (*Zafarani, Abbasi & Liu, 2014*). Precision is important when the cost of a false positive is high. Applying symmetrical logic, recall would be important when the cost of a false negative is high. When identifying tweets with anti-government sentiment, a false positive implies that a non anti-government tweet is labeled as anti-government, whereas a false negative implies that an anti-government tweet is labeled as non anti-government. In false positive labeling, because a non anti-government tweet may be labeled as anti-government, it may be subject to one or more stringent misinformation policies such as being tagged with a warning label or turning off the likes and retweets to curb their spread, or ultimately, demoting or removing the tweet altogether (*Twitter, 2021*). These measures may raise allegations of freedom of speech violations (*Parler, 2022*). On the other hand, false negative labeling implies that an anti-government tweet will be labeled as non anti-government. This tweet may propagate across the network unhindered, spreading anti-government agenda. However, it will steer clear of freedom of expression violations. Absent a clear threat, where there may be a compelling reason, such as an explosive political environment, to curb the spread of anti-government tweets, a balance may be sought between precision and recall to trade-off between the diffusion of anti-government sentiment and violating freedom of expression. F1-score provides this balance between the two metrics.

RESULTS AND DISCUSSION

We split the entire *corpus* using stratified sampling into two partitions; training and test consisting of 80% and 20% of the tweets, respectively. Stratified sampling preserves the ratio of anti-government to non anti-government tweets in each partition. With each split, we conducted extensive experimentation with three levels of feature engineering. Each level of feature engineering draws upon the results from the previous level, and is designed to allow for an increasing selectivity of features.

The performance of the classifiers for the three experiments is summarized in Table 7. The table also includes the time taken to train each model. The first experiment included all features, except for those interaction and authors' features which were not statistically significant. Collectively, this set consisted of 8,203 features, and they identified anti-government tweets with an accuracy of 86% and F1-score of 0.81. All the models, except for decision trees, offered competitive performance. It was not surprising that the performance of decision trees was significantly lower because these simple learners are prone to over-fitting (*Liaw & Wiener, 2002*). It was perhaps more surprising that the simple logistic regression model came close to complex models such as the multi-layer perceptron and support vector machines. The superior performance of the logistic regression model could be because much of the classification decision was based on the textual content of the tweets, similar to the detection of hate speech (*Khan et al., 2021*), a conjecture that we confirmed through importance analysis.

Figure 4 shows the importance scores of the various groups of features computed using the random forest model. Guided by the feature map in Fig. 1, we further grouped features into coarser categories. In Fig. 4, social features include those extracted from the structured

Table 7 Performance metrics.

Expt.	Model	Accuracy	Precision	Recall	F1-score	Training time
I	Linear SVC	0.82	0.76	0.77	0.76	3 min 5 s
	Support vector	0.85	0.75	0.82	0.77	10 min 44 s
	Decision tree	0.72	0.64	0.64	0.64	9.43 s
	Random forest	0.84	0.69	0.85	0.73	19.3 s
	Logistic regression	0.84	0.78	0.78	0.78	10.7 s
	Multi-layer perceptron	0.86	0.80	0.82	0.81	2 min 14 s
	DistilBERT	0.84	0.82	0.76	0.78	1 min 43 s
II	Linear SVC	0.83	0.79	0.77	0.78	1 min 43 s
	Support vector	0.85	0.80	0.80	0.80	13.4 s
	Decision tree	0.71	0.62	0.61	0.62	1.52 s
	Random forest	0.83	0.69	0.84	0.72	5.47 s
	Logistic regression	0.81	0.79	0.75	0.77	0.53 s
	Multi-layer perceptron	0.85	0.78	0.8	0.79	15.8 s
III	Linear SVC	0.83	0.80	0.77	0.78	6.01 s
	Support vector	0.87	0.82	0.83	0.82	10.5 s
	Decision tree	0.74	0.64	0.65	0.65	4.2 s
	Random forest	0.82	0.67	0.83	0.70	16.8 s
	Logistic regression	0.82	0.80	0.76	0.77	0.33 s
	Multi-layer perceptron	0.86	0.78	0.82	0.80	8.79 s

data related to the tweets and their authors. These comprise of the interaction parameters of the tweets, and the network strength, activity level, and the authenticity status of the authors. Auxiliary features consists of punctuation counts, VADER sentiment, emotion, and readability scores. The figure shows that TF-IDF scores extracted from the text of the tweets and the profile descriptions of the authors contribute about 74% to the classification. It was thus possible to hypothesize that only a small subset of the features would be sufficient to achieve the same classification accuracy. Therefore, in the second experiment, we selected the top 300 features and re-trained and re-evaluated the classifiers. The metrics for the different classifiers after employing feature selection indicate that this step does not improve the performance of the classifiers. However, the table shows an appreciable reduction in the training time by employing feature selection. In fact, the training time of the support vector classifier, which is the model that offers the best performance, reduced from 10 min 44 s when the entire collection of features is used to merely 13.4 s when only the top 300 features were selected.

In the third experiment, we reduced the dimensionality of the top 300 features chosen in the second experiment *via* principal components analysis (*Jolliffe & Cadima, 2016*). The performance of support vector classifier with RBF kernel increases by 1%, and the time to train drops slightly to 10.5 s. Thus, feature selection and dimensionality reduction together offer a distinct advantage of improving the efficiency of the classification by reducing the number of features without sacrificing performance. Moreover, precision and recall

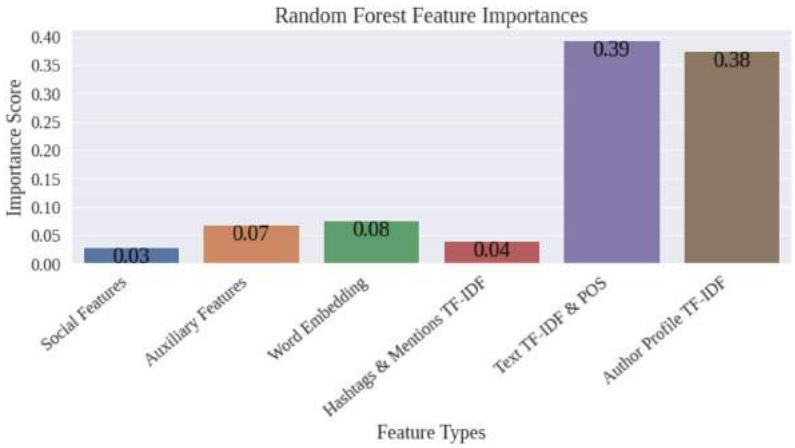


Figure 4 **Feature selection—importance scores.** Full-size DOI: 10.7717/peerj-cs.1127/fig-4

metrics are higher and more balanced, leading to a better F1-score when feature processing is employed than when it is not.

The transformer model DistilBERT is trained using the text from the tweets and authors' profiles, along with hashtags. Interaction and authors' features are not included in training the transformer model. Moreover, feature engineering cannot be applied to DistilBERT, and hence, the results of DistilBERT are compared with the performance metrics of the models from the first experiment in the table. The table shows that even when feature engineering is not employed for conventional machine learning models, DistilBERT is not the best performing model. In fact, it is outperformed by multi-layer perceptron. With subsequent feature engineering, DistilBERT is outperformed by both support vector and multi-layer perceptron classifiers. Thus, although pre-trained transformer models represent the state-of-the-art in natural language processing (*Nagda et al., 2020*), extensively trained conventional machine learning models with additional features extracted from meta data and careful feature selection and dimensionality reduction can outperform these models.

Our objective in this article was to build a classification framework that can detect deviant content that promotes an anti-government perspective. The basis of this framework is a set of features extracted from the structured and unstructured tweet data that we expect to remain invariant in conversations on various controversial topics and issues. Using these features, accompanied by feature selection and processing along with tuning of hyperparameters of machine learning models, we have also successfully applied this framework to detect tweets that spread anti-mask (*Cerbin et al., 2021*) and anti-vaccination (*Paul & Gokhale, 2020*), and those that support Proud Boys, an extremist, radical group (*Fahim & Gokhale, 2021*).

CONCLUSIONS AND FUTURE RESEARCH

Radical extremists use social media platforms to spread their ideology effectively in order to gather a critical mass of followers to organize and execute violent and disruptive activities in physical spaces. Such anti-government sentiment can simmer on social media

platforms for a while, and can be a harbinger of disruption, bloodshed, and unrest downstream. Analyzing social media dialogue can detect these latent views and stop them for escalating further. This article presented a classification framework that detects anti-government sentiment in social media dialogue following the anti-lockdown protests in Lansing, Michigan during the COVID-19 pandemic. Using the tweets collected and labeled from two separate protests, we computed a rich set of features from both structured and unstructured data returned by Twitter. These features were processed using feature selection and dimensionality reduction, and were then used to train popular machine learning models. Our framework could efficiently separate anti-government sentiment with approximately 87% accuracy, balancing precision and recall (F1-score of 0.82), and with a training time of only a few seconds. This anti-government propaganda was immersed in various contextual and circumstantial information, hence, lacked clear focus or philosophy. The research thus demonstrated the promise of feature engineering and machine learning to detect deviant content that can precipitate violence even though it is submerged in the surrounding events. It thus opens up the possibility of employing these techniques to identify and demote deviant chatter on social media platforms before it can cause offline damage, as well as gateways for future advances on this topic.

Our future research involves building methods to geo-locate tweets to understand the geographical dispersion of anti-government, extremist content.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Hieu Nguyen conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Swapna Gokhale conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub: https://github.com/HieuQN/Michigan-Protest.

REFERENCES

Abd-Elaal AIA, Badr AZ, Mahdi H. 2020. Detecting violent radical accounts on Twitter. *International Journal of Advanced Computer Science and Applications* **11(8)**:516–522 DOI 10.14569/issn.2156-5570.

Ahmad S, Asghar MZ, Alotaibi FM, Awan I. 2019. Detection and classification of social media-based extremist affiliation using sentiment analysis techniques. *Human-Centric Computing and Information Sciences* **9**:24 DOI 10.1186/s13673-019-0185-6.

Aliapoulios M, Bevensee E, Blackburn J, Bradlyn B, Cristofaro ED, Stringhini G, Zannettou S. 2021. An early look at the Parler online social network. *International AAAI Conference on Web and Social Media* **15**:943–951 DOI 10.48550/arXiv.2101.03820.

Andone D. 2020. Protests are popping up across the US over stay-at-home restrictions. *Available at https://www.cnn.com/2020/04/16/us/protests-coronavirus-stay-home-orders/index.html* (accessed 5 July 2022).

Araque O, Iglesias CA. 2020. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access* **8**:17877–17891 DOI 10.1109/ACCESS.2020.2967219.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A. 2018. Quanteda: an R package for the quantitative analysis of textual data. *Journal of Open Source Software* **3(30)**:774 DOI 10.21105/joss.00774.

Berg K, Egan P. 2020. Thousands converge to protest Michigan governor's stay-home order in Operation Gridlock. *Available at https://www.usatoday.com/story/news/nation/2020/04/15/lansing-capitol-protest-michigan-stay-home-order/5139472002/* (accessed 5 July 2022).

Bogel-Burroughs N. 2020. What we know about the alleged plot to kidnap Michigan's governor. *Available at https://www.nytimes.com/2020/10/09/us/michigan-militia-whitmer.html*.

Caldarelli G, De Nicola R, Vigna F, Petrochi M, Saraco F. 2020. The role of bot squads in political propaganda on Twitter. *Communications Physics* **3**:81 DOI 10.1038/s42005-020-0340-4.

Cerbin L, DeJesus J, Warnken J, Gokhale SS. 2021. Unmasking the mask debate on social media. In: *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. Piscataway: IEEE, 677–682.

Clarke CP. 2022. Op-Ed: the newest variant of violent extremism? Using paranoia about the pandemic as a recruiting tool. *Available at https://www.latimes.com/opinion/story/2022-01-09/covid-vaccines-paranoia-recruiting-extremists-terrorism*.

Cotfas L-A, Delcea C, Gherai R, Roxin I. 2021. Unmasking people's opinions behind mask-wearing during COVID-19 pandemic: a twitter stance analysis. *Symmetry* **13(11)**:1995 DOI 10.3390/sym13111995.

DeCook JR. 2018. Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology* **43**:485–504 DOI 10.1080/17439884.2018.1544149.

Delashmit WH, Manry MT. 2005. Recent developments in multilayer perceptron neural networks. In: *7th Annual Memphis Area Engineering and Science Conference*.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint*. DOI 10.48550/arXiv.1810.04805.

Ecarma C. 2020. Trump supporters are staging armed protests to stick it to Coronavirus. *Available at https://www.vanityfair.com/news/2020/04/trump-supporters-protest-coronavirus-orders* (accessed 5 July 2022).

Fahim M, Gokhale SS. 2021. Identifying social media content supporting proud boys. In: *2021 IEEE International Conference on Big Data*. Piscataway: IEEE, 2487–2495.

Grill G. 2021. Future protest made risky: examining social media based civil unrest prediction research and products. *Computer Supported Coop Work* **30**:811–839 DOI 10.1007/s10606-021-09409-0.

Hajibagheri A, Sukhthankar G. 2014. Political polarization over global warming: analyzing Twitter data on climate change (poster). In: *ASE International Conference on Social Computing*. Palo Alto, CA.

Helliwell JF, Huang H, Wang S, Norton M. 2021. Happiness, trust and deaths under COVID-19. *Available at https://worldhappiness.report/ed/2021/happiness-trust-and-deaths-under-covid-19/* (accessed 31 January 2022).

Hernandez S. 2020. This is how a group linked to Betsy DeVos is organizing protests to end social distancing, now with Trump's support. *Available at https://www.buzzfeednews.com/article/salvadorhernandez/coronavirus-quarantine-protests-facebook-groups* (accessed 5 July 2022).

Hutto C, Gilbert E. 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 216–225.

Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **374(2065)**:20150202 DOI 10.1098/rsta.2015.0202.

Kearney MW, Sancho RL, Wickham H, Heiss A, Briatte F, Sidi J. 2020. Collecting Twitter data. *Available at https://cran.r-project.org/web/packages/rtweet/rtweet.pdf*.

Khan H, Yu F, Sinha A, Gokhale SS. 2021. A parsimonious and practical approach to detecting offensive speech. In: *2021 International Conference on Computing, Communication, and Intelligent Systems*. 688–695.

Klein A. 2019. From Twitter to Charlottesville: analyzing the fighting words between the alt-right and antifa. *International Journal of Communication* **13**:297–318.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2(3)**:18–22.

Loper E, Bird S. 2002. NLTK: the natural language toolkit. *CoRR* DOI 10.48550/arXiv.cs/0205028.

Mann E, Gaines B, Gokhale S. 2022. An analysis of the early dialogue on vaccine passports on Twitter. In: *2022 IEEE 45th Computer Software and Applications Conference*. Piscataway: IEEE.

Mauger C. 2020. Protesters, some armed, enter Michigan Capitol in rally against COVID-19 limits. *Available at https://www.detroitnews.com/story/news/local/michigan/2020/04/30/protesters-gathering-outside-capitol-amid-covid-19-restrictions/3054911001/* (accessed 5 July 2022).

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In: *Neural Information Processing Systems*. Vol. 2. 3111–3119.

Miranda E, Aryuni M, Fernando Y, Kibtiah TM. 2020. A study of radicalism contents detection in Twitter: insights from support vector machine technique. In: *2020 International Conference on Information Management and Technology*. 549–554.

Mussiraliyeva S, Bolatbek M, Omarov B, Medetbek Z, Baispay G, Ospanov R. 2020. On detecting online radicalization and extremism using natural language processing. In: *2020 21st International Arab Conference on Information Technology (ACIT)*. 1–5.

Nagda K, Mukherjee A, Shah M, Mulchandani P, Kurup L. 2020. Ascent of pre-trained state-of-the-art language models. In: Vasudevan H, Michalas A, Shekokar N, Narvekar M, eds. *Advanced Computing Technologies and Applications*. Singapore: Springer, 269–280.

Nguyen L, Othmeni O. 2021. The rise of digital extremism: how social media eroded America's political stability. *Available at https://www.ivint.org/the-rise-of-digital-extremism-how-social-media-eroded-americas-political-stability/* (accessed 21 January 2020).

Nicholson C. 2019. A beginner's guide to word2vec and neural word embeddings. *Available at https://pathmind.com/wiki/word2vec*.

Oehmichen A, Hua K, Amador Diaz Lopez J, Molina-Solana M, Gomez-Romero J, Guo Y-K. 2019. Not all lies are equal. a study into the engineering of political misinformation in the 2016 US presidential election. *IEEE Access* **7**:126305–126314.

Park PS, Compton RF, Lu T-C. 2015. Network-based group account classification. In: *Eighth International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*. 163–172.

Parler. 2022. Parler—where free speech thrives. *Available at https://parler.com/* (accessed 21 January 2020).

Paul N, Gokhale SS. 2020. Analysis and classification of vaccine dialogue in the coronavirus era. In: *2020 IEEE International Conference on Big Data (Big Data)*. Piscataway: IEEE, 3220–3227.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12(Oct)**:2825–2830 DOI 10.48550/arXiv.1201.0490.

Qi H, Jiang H, Bu W, Zhang C, Shim KJ. 2019. Tracking political events in social media: a case study of Hong Kong protests. In: *2019 IEEE International Conference on Big Data*. Piscataway: IEEE, 6192–6194.

Rehurek R, Sojka P. 2010. Software framework for topic modelling with large corpora. In: *LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.

Reid SE, Valasik M, Bagavati A. 2020. Examining the physical manifestation of alt-right gangs: from online trolling to steet fighting. In: Melde C, Weerman F, eds. *Gangs in the Era of Internet and Social Media*. Berlin: Springer, 105–134.

Sanh V, Debut L, Chaoumond J, Wolf T. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint* DOI 10.48550/arXiv.1910.01108.

Saravia E, Liu H-C, Huang Y-H, Wu J, Chen Y-S. 2018. CARER: contextualized affect representations for emotion recognition. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3687–3697.

Skelton M. 2022. The world happiness report 2022—happiness is about benevolence and trust. *Available at https://marieskelton.com/happiness-is-about-benevolence-and-trust/* (accessed 30 April 2022).

Stukal D, Sanovich S, Bonneau R, Tucker JA. 2022. Why botter: how pro-government bots fight opposition in Russia. *American Political Science Review* **116(3)**:1–15 DOI 10.1017/S0003055421001507.

Suykens JAK, Lukas L, Vandewalle J. 2000. Sparse least squares support vector machine classifiers. In: *Neural Processing Letters*. 293–300.

Talburt J. 1985. The Flesch index: an easily programmable readability analysis algorithm. In: *Proceedings of SIGDOC*. 114–122.

Tweettabs. 2022. Like, retweet and quote tweet: understanding the Twitterverse. *Available at https://www.tweettabs.com/how-to-quote-a-tweet/* (accessed 31 January 2022).

Twitter. 2021. COVID-19 misleading information policy. *Available at https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy* (accessed 31 January 2022).

UN. 2012. The use of the internet for terrorist purposes. *Available at https://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf* (accessed 31 January 2022).

Willett P. 2006. The Porter stemming algorithm: then and now. *Program Electronic Library and Information Systems* **40(3)**:219–223 DOI 10.1108/00330330610681295.

Wilson J. 2020. The rightwing groups behind wave of protests against COVID-19 restrictions. *Available at https://www.theguardian.com/world/2020/apr/17/far-right-coronavirus-protests-restrictions* (accessed 5 July 2022).

Wolfowicz M, Perry S, Hasisi B, Weisburd D. 2021. Faces of radicalism: differentiating between violent and non-violent radicals by their social media profiles. *Computers in Human Behavior* **116(3)**:106646 DOI 10.1016/j.chb.2020.106646.

Wu C, Gerber MS. 2018. Forecasting civil unrest using social media and protest participation theory. *IEEE Transactions on Computational Social Systems* **5(1)**:82–94 DOI 10.1109/TCSS.2017.2763128.

Xu R. 2014. Pos weighted tf-idf algorithm and its application for an mooc search engine. In: *2014 International Conference on Audio, Language and Image Processing*. 868–873.

Yasin R, Lutfi S, Imene A, Oroumchian F, el Barachi M, Mathew SS. 2021. Study of radical views on social media: classification and group dynamics analysis. In: *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*. 1–8.

Zafarani R, Abbasi MA, Liu H. 2014. *Social media mining: an introduction*. Cambridge: Cambridge University Press.

PeerJ Computer Science

Sentiment analysis of vegan related tweets using mutual information for feature selection

Elvina Shamoi[1], Akniyet Turdybay[1], Pakizar Shamoi[1], Iskander Akhmetov[1,2], Assel Jaxylykova[2,3] and Alexandr Pak[1,2]

[1] School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan
[2] Institute of Information and Computational Technologies, Almaty, Kazakhstan
[3] Kazakh National University, Almaty, Kazakhstan

ABSTRACT

Nowadays, people get increasingly attached to social media to connect with other people, to study, and to work. The presented article uses Twitter posts to better understand public opinion regarding the vegan (plant-based) diet that has traditionally been portrayed negatively on social media. However, in recent years, studies on health benefits, COVID-19, and global warming have increased the awareness of plant-based diets. The study employs a dataset derived from a collection of vegan-related tweets and uses a sentiment analysis technique for identifying the emotions represented in them. The purpose of sentiment analysis is to determine whether a piece of text (tweet in our case) conveys a negative or positive viewpoint. We use the mutual information approach to perform feature selection in this study. We chose this method because it is suitable for mining the complicated features from vegan tweets and extracting users' feelings and emotions. The results revealed that the vegan diet is becoming more popular and is currently framed more positively than in previous years. However, the emotions of fear were mostly strong throughout the period, which is in sharp contrast to other types of emotions. Our findings place new information in the public domain, which has significant implications. The article provides evidence that the vegan trend is growing and new insights into the key emotions associated with this growth from 2010 to 2022. By gaining a deeper understanding of the public perception of veganism, medical experts can create appropriate health programs and encourage more people to stick to a healthy vegan diet. These results can be used to devise appropriate government action plans to promote healthy veganism and reduce the associated emotion of fear.

Subjects Artificial Intelligence, Computational Linguistics, Network Science and Online Social Networks, Text Mining, Sentiment Analysis
Keywords Natural language processing, Sentiment analysis, Emotion detection, Social media monitoring, Opinion mining, Twitter sentiment classification, Vegan tweets, Pointwise mutual information, Sentiment tracking

Submitted 10 June 2022
Accepted 17 October 2022
Published 5 December 2022

Corresponding authors
Pakizar Shamoi, p.shamoi@kbtu.kz
Iskander Akhmetov,
iskander.akhmetov@gmail.com

Academic editor
Ana Reyes-Menendez

Additional Information and Declarations can be found on page 32

DOI **10.7717/peerj-cs.1149**

Copyright
2022 Shamoi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

A vegan diet (or plant-based diet) assumes refraining from consuming animal or animal-derived products, such as meat, fish, dairy, eggs, cheese, *etc.*, (*MacInnis & Hodson, 2017*). Fruits, vegetables, beans, whole grains, legumes, nuts, and a range of other whole plant

foods make up this diet. The motivations for switching to a plant-based diet usually fall into one of the three categories: health benefits, animal rights, and environmental protection.

Vegans have traditionally been portrayed negatively in society and social media. There is even a term to describe this opposition: *vegaphobia* (*Cole & Morgan, 2011*). Vegaphobic sentiments and bias towards vegans appeared in the 2010s (*MacInnis & Hodson, 2017*).

However, in recent years, the COVID-19 pandemic and global warming have increased the awareness of plant-based lifestyles and indeed produced significant changes in people's daily lives (*Park & Kim, 2022*; *WHO, 2021*; *Rzymski et al., 2021*; *Alae-Carew et al., 2022*). Moreover, in 2015, based on numerous studies, the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO) declared that processed meat is a group 1 carcinogen and that red meat is a group 2 carcinogen (*Willett et al., 2019*). At the same time, several recent studies (*Dagevos & Voordouw, 2013*; *Willett et al., 2019*) report that a plant-based diet is better for the environment and more sustainable than other kinds of diet. Because of all those reasons, combined with benefits for human health and environmental sustainability, interest in a plant-based diet is expanding. These days of climate change, this subject is becoming increasingly significant and is frequently debated on social media.

Besides sustainability issues, one of the key motives for following a vegan diet is for health reasons (*Radnitz, Beezhold & DiMatteo, 2015*; *Jennings et al., 2019*). A vegan diet has been linked in several studies with better health and longevity, reduced blood pressure and cholesterol, lower rates of heart disease, type 2 diabetes, and various forms of cancer (*Willett et al., 2019*; *Norman & Klaus, 2019*; *Qian et al., 2019*). Globally, there is a rise in the number of people with obesity, which is associated with many serious medical problems (*Finucane et al., 2011*; *Revels, Kumar & Ben-Assuli, 2017*). Moreover, research indicates that obesity is linked to a higher risk of severe COVID-19, cardiovascular disease, and cancer (*Finucane et al., 2011*). At the same time, studies prove that a plant-based diet helps individuals to lose weight and prevent obesity (*Tran et al., 2020*; *Turner-McGrievy, Mandes & Crimarco, 2017*).

Several studies aimed to find the association between diet and eating habits with the risk and severity of COVID-19 (*Aman & Masood, 2020*; *James et al., 2021*; *Zabetakis, Matthys & Tsoupras, 2021*; *Kim et al., 2021*; *Merino et al., 2021*). Plant-based diets were linked to a lower risk of moderate-to-severe COVID-19 (*Kim et al., 2021*). Another study revealed that a diet rich in plant-based foods had been linked to reduced risk and severity of COVID-19 (*Merino et al., 2021*). Therefore, a vegan diet may offer protection against severe COVID-19.

Another study (*Di Renzo et al., 2020*) reports that during the COVID-19 pandemic, population lifestyle and eating habits changed. Specifically, weight gain was observed in 48.6% of the respondents; 3.3% decided to stop smoking; 38.3% reported an increase in physical activity; 15% of the participants started purchasing organic fruits and vegetables, and many people aged 18-30 years shifted to the Mediterranean diet. These findings are also supported by another study (*Singh, Jain & Rastogi, 2021*), which reported a substantial decrease in consumption of junk food (73.8%), alcohol (27.6%), and smoking (8.1%).

Previous research has revealed that the number of people following a vegan or vegetarian diet is increasing (*López et al., 2019*). Another study showed that individuals regard vegan, organic, and homemade foods as the healthiest options (*Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*). The three basic characteristics of food that were most communicated in 2020 were vegan, homemade, and organic.

Many people's daily lives are now dominated by social media (*Ostic et al., 2021*). People use social networks to connect with other people, share opinions on various subjects, study and work. Understanding people's views, experiences, and attitudes through social media analysis can play a vital role in government action program development (*Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*). In turn, this is crucially important in promoting healthy eating, tackling global warming, and lowering a person's COVID-19 risk, among others. One of the social media to express people's opinions is Twitter. Twitter is one of the most widely used social media (*Williams, Terras & Warwick, 2013*) that allows users to write everything that comes to their mind in the form of tweets (*Williams, Terras & Warwick, 2013*).

The present article aims to utilize Twitter posts to understand public opinion on social media better as expressed on Twitter regarding the vegan (plant-based) diet from 2010 to 2022. We also plan to identify what was the impact of the pandemic on adherence to a vegan diet specifically and the main topics (words) when veganism is discussed in a positive or negative context. The study employs a data set containing a collection of popular vegan-related tweets.

As we see, previous studies on plant-based diets were primarily survey-based and conducted from April to May 2020. On the other hand, we use Twitter data, not surveys, to examine how public perception of veganism changed over a 12-year period, whether people felt positive or negative about vegan diets, and the prevailing emotions behind those sentiments.

We seek to answer the following research questions using vegan tweets from 2010–2012:

- How has the public perception of veganism changed over the past 12 years?
- What were the prevailing emotions associated with this change?
- Did the global pandemic of COVID-19, climate crisis, vegan documentaries, and research unveiling the health benefits of the vegan diet affect the attitude toward the vegan diet on Twitter?

The article presents the first comprehensive analysis of the public perception of the vegan diet from 2010 to 2022 on Twitter with an emphasis on new insights into the key emotions associated with this growing trend. Association with the key factors influencing this trend throughout the period was also done. Considering that:

- In 2018, UN IPCC reported that climate change catastrophe would occur in the upcoming 12 years (*Masson-Delmotte et al., 2019*);
- A plant-based diet is more sustainable and better for the environment than other diets, according to recent studies (*Dagevos & Voordouw, 2013*; *Willett et al., 2019*; *Scarborough et al., 2014*). For example, by comparing beef and soybeans, one study investigated that

soy needs 18 times less energy per gram of protein and produces 71 times less CO_2 than beef (*Cleveland & Gee, 2017*). Another study indicates that the vegan diet has the lowest CO_2 emission compared to other diets (*Chai et al., 2019*). An Oxford University study reports that meat-eaters are responsible for about two and a half times as much dietary greenhouse gas emissions per day as vegans (*Scarborough et al., 2014*);

- Environmental sustainability in these days of climate change is of primary importance (*Masson-Delmotte et al., 2019*; *Dagevos & Voordouw, 2013*),

this topic is of vital significance. By gaining a better understanding of the opinion on veganism and associated emotions, medical professionals can design effective health initiatives and encourage more people to either stick to a healthy plant-based diet or reduce animal product consumption. Even small dairy and meat reductions in many people may make a valuable contribution to climate change (*Scarborough et al., 2014*). According to recent research on the quantitative estimate of the potential climate effects of a hypothetical transition to a plant-based diet worldwide, such a phaseout can stabilize greenhouse gas levels for 30 years and offset 68% of CO_2 emissions this century *Eisen & Brown (2022)* (see Fig. 1).

We aim to use sentiment analysis to track the change in public perception of veganism. Sentiment analysis is a computational study for identifying emotions represented in texts. The purpose of sentiment analysis is to classify the texts' polarity—determine whether a piece of text (tweet in our case) conveys a negative or positive viewpoint (*Neethu & Rajasree, 2013*; *Chinnalagu & Durairaj, 2021*). Therefore, sentiment analysis is mainly used to grasp opinions about trends, products, services, or political issues.

Although there are various approaches to constructing sentiment analysis models, we chose the point-wise mutual information (PMI) approach proposed by *Church & Hanks (2002)*, *Turney & Littman (2003a)*, *Turney & Littman (2003b)* and *Turney (2002)* for its interpretability and robustness to statistical bias in small sample sizes. Another motivation for using the mentioned approach is its application to perform feature selection in this study. MIT researcher Robert Fano introduced the basics of the approach concept in 1961 (*Fano, 1961*) and *Jurafsky et al. (2014)* initially referred to as mutual information (MI). The PMI or MI in NLP application measures the probability of two-word co-occurrence relative to the random probability, thus giving more sense to the semantic proximity of the words. Therefore, our rationale behind using the method is to see how effectively the words sought to be associated with certain sentiment classes (PMI measures) can serve as features in training the sentiment classifier model. A number of studies use the PMI-based approach to sentiment analysis (*Bindal & Chatterjee, 2016*; *Vo & Ock, 2012*; *Feldman, 2013*; *Hamdan, Bellot & Bechet, 2015*; *Utama, 2019*; *Bonzanini, 2016*; *Kanna & Pandiaraja, 2019*).

The article is structured as follows. Section I is this introduction. An overview of the sentiment analysis techniques is presented in Section II. In Section III, we describe our method—the Pointwise Mutual Information-Based Feature Selection for Sentiment Analysis. In addition, we perform emotion analysis at this stage to take a more in-depth look at the public's interest in veganism. Data collection and data annotation procedures are also described there. Evaluation of our method along with experimental results is shown

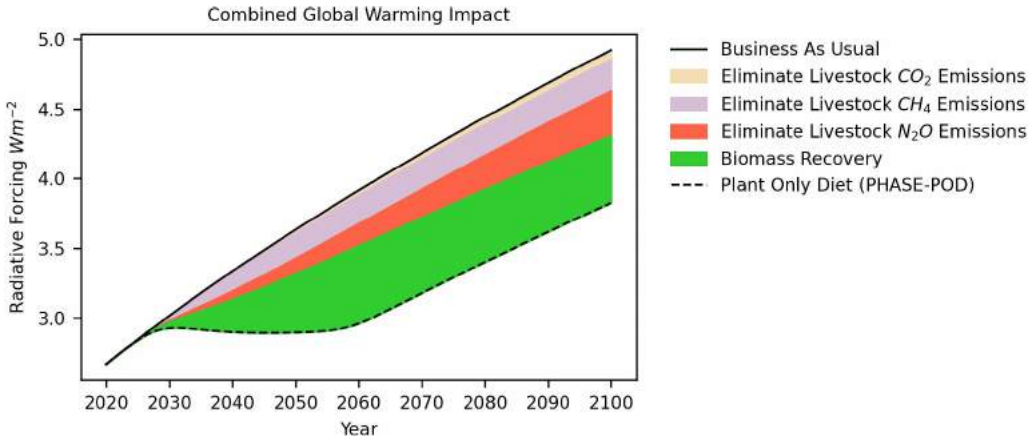


Figure 1 Global phaseout of animal agriculture and transition to a plant-based diet can potentially reduce global warming's impact on the atmosphere. Reprinted from *Eisen & Brown (2022)*.
Full-size DOI: 10.7717/peerjcs.1149/fig-1

in Section IV. Finally, concluding remarks are provided in Section V. The last section also presents insights on how the method can be improved in the future.

The following section provides an overview of sentiment analysis techniques.

SENTIMENT ANALYSIS: THEORETICAL BACKGROUND

Approaches for sentiment analysis

The unprecedented growth of user-generated content that we are witnessing nowadays on the World Wide Web (WWW) implies that for the first time in humankind's history, the broad public can express their opinions on various subjects almost effortlessly. Of course, this happened with the help of contemporary technologies and Social Network (SN) phenomena emerging in the early 2000s. And analyzing public opinions on subjects of interest is vital for government and business.

Sentiment analysis (SA) or opinion mining (OM) is identifying and classifying people's attitudes toward a subject or entity using the quantitative methods of statistics, computational linguistics, and machine learning (ML) (*Neethu & Rajasree, 2013*). The sentiment analysis topic has been a subject of rigorous research in the last years (*Ligthart, Catal & Tekinerdogan, 2021*), and the beginning of the research in this area can be traced back to the middle of the 20th century (*Sims, 2015*).

SA tasks can be divided into the following major categories (*Hemmatian & Sohrabi, 2019*; *Ligthart, Catal & Tekinerdogan, 2021*):

1. **Opinion extraction** is the process of user opinion extraction from WWW to discover their way of thinking.
2. **Sentiment mining** is the separation of objective or factual sentences from the subjective ones that contain opinions and sentiments.
   Subjectivity classification attributes a sentence to a class of subjectivity it carries.
3. **Sentiment classification** categorizes the opinions discovered into predefined classes.
   Polarity determination whether it is positive, negative, or lacks the polarity (neutral).

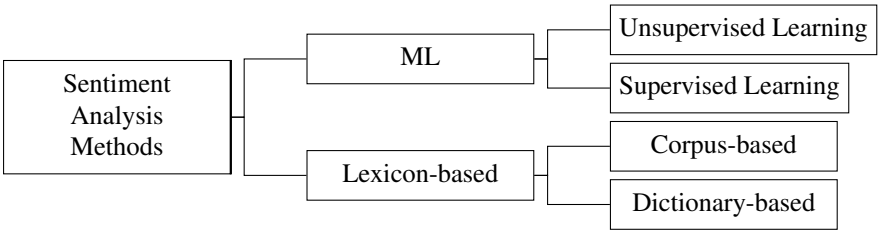


Figure 2 **Sentiment analysis methods classification.**
Full-size DOI: 10.7717/peerjcs.1149/fig-2

The approaches for sentiment analysis can be classified as follows (*Adwan et al., 2020*); see also Fig. 2:

1. **Lexicon based** approaches use dictionaries (*Hu & Liu, 2004*) of words that are assumed to be positive, negative, or neutral, and then weigh or score the text according to the relative content of these words to infer its sentiment (*Akilandeswari & Jothi, 2018*; *Das et al., 2018*; *Azizan et al., 2019*).
   Dictionary-based approaches use a vocabulary look-up technique to detect sentiment-rich words and their polarity.
   Corpus-based approaches work with the statistical models developed on many texts to leverage the information on the context of word use and the dictionary-based approach.
2. **ML** approaches can be subdivided into Supervised and Unsupervised Learning classes of tasks:
   Supervised learning methods require a dataset with items labeled for the sentiment to train classification modeling algorithms (SVM, KNN, DTC, LR, and others).
   Unsupervised learning methods search for the structure of information within the presented data using clustering, topic modeling, and mapping algorithms (*Mukhamediev et al., 2020*; *Kirill et al., 2020*).
3. **Hybrid** approaches, as the name implies, use both the Lexicon-based and ML approaches together.

A challenging task in sentiment analysis is the domain specificity of the developed approaches, which come from the fact that sentiment-rich words can carry different polarity in usage across domains. For example, the term "*aggressive*" can be perceived as positive when related to sports ("*Chelsea played aggressively tonight in a match with Barcelona club.*") and negative in the political domain ("*Russian troops aggressively attack Mariupol city in Ukraine, thousands of civilians get injured.*"). Thus, topic-aware or domain-specific models are advantageous over the general ones (*Akhmetov, Gelbukh & Mussabayev, 2022*).

Another problem that definitely should be addressed is the possible data imbalance on the sentiment classes, especially in the domains where most publications are negative (war, accident, crime news). Handling the data imbalance can be done with under-over-sampling techniques, data augmentation, and others (*Omarkhan, Kissymova & Akhmetov, 2021*; *Kelsingazin, Akhmetov & Pak, 2021*).

One of the classical works on the Twitter micro-blog post (tweet) analysis is VADER (*Hutto & Gilbert, 2014*), a simple rule-based sentiment analysis model. VADER employs

an empirically validated gold-standard lexical feature list with the associated sentiment or opinion polarity and specializes in micro-blog short text messages. The model shows an F1 classification accuracy of 0.96, outperforming human raters with a 0.84 result. Another model trained on Twitter data (*Kocaman & Talby, 2021*) shows an accuracy of 0.80.

Related works

Analyzing posts in SN can be of particular interest to government administration, business officials from the marketing and sales side, social scientists, and anthropologists, as they contain vast amounts of information characterizing the post writers' origins, cultural background, political views, and attitudes to a broad spectrum of objects, events and more abstract things as ideology.

One of the most popular SN is Twitter: the micro-blogging platform, where its 1.3 billion users make 500 million tweets every day (*Alhgren, 2022*). Analyzing Twitter posts or the so-called tweets involves exploring the interest, thoughts, and opinions in various aspects and domains (*Adwan et al., 2020*).

The application examples of tweet analysis are to understand people's opinions towards different topics such as COVID-19, disasters, presidential elections, armed conflicts across the Globe (Ukraine, Syria, Palestine, Iraq), sports, and healthy lifestyles. Much research has focused on mining public opinion on various social, cultural, economic, and environmental topics on Twitter (*Reyes-Menendez, Saura & Alvarez-Alonso, 2018*; *Reyes-Menendez, Saura & Filipe, 2020*; *Saura, Reyes-Menendez & Palos-Sanchez, 2019*; *Serrano et al., 2021*; *Das et al., 2018*; *Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*; *Islam, 2019*; *Gore, 2015*; *Karami et al., 2018*; *Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*; *Saura, Palos-Sanchez & Rios Martin, 2018*). *Reyes-Menendez, Saura & Alvarez-Alonso (2018)* investigated the main factors that worry people about the environment and public health. Analysis of health-related tweets by *Islam (2019)* revealed an interesting correlation between yoga and veganism. *Gore (2015)* studied geographic issues of obesity rate in America. The main characteristics of public perceptions of obesity, diabetes, nutrition, and exercise were explored by *Karami et al. (2018)*. *Reyes-Menendez, Saura & Filipe (2020)* identified the main challenges involved in the *MeToo* movement. In *Saura, Reyes-Menendez & Palos-Sanchez (2019)* authors explored negative and positive perceptions of consumers related to the Black Friday event. Next, *Serrano et al. (2021)* examines online reviews of green Airbnb users to find the main aspects among which the term *sustainability* appears. *Pila, Kvasnikov Stanislavsk & Kvasnika (2021)* explored all tweets between 2019 and 2020 that included the hashtag "*healthyfood*".

METHODS

Data collection

We study Twitter messages or tweets related to veganism to gain insight into the change in the sentiment of such posts during the COVID-19 world pandemic, before it, and just after it, including publication of studies and documentaries about veganism.

Kaggle dataset

For Sentiment analysis classification model training, we used the Sentiment140 dataset (https://www.kaggle.com/datasets/kazanova/sentiment140) with 1.6 million year 2009 tweets labeled for sentiment (0 = negative, and 4 = positive). We filtered the dataset by selecting only tweets containing the following keywords: "*vegan*", "*plantbased*", "*healthy*", "*vegetarian*", "*veggie*", "*veganism*", "*cruelty free*", "*plant milk*", "*beyond meat*".

After filtering the tweets by vegan keywords, we ended up with a dataset of 1,876 tweets containing our target keywords. The initial dataset contained equal amounts of positive and negative labels or 800K each. However, the filtered tweets include 798 negative and 1,078 positive tweets. Thus, initially, the *vegan tweets* are more positive: the mean sentiment is 2.3 as opposed to the general score of 2.0 for all the tweets in the original dataset.

The most popular vegan keyword is "*healthy*" occurring in 955 of the selected tweets, followed by "*veggie*" (391), "*vegan*" (364) and "*vegetarian*" (191). For some of the keywords ("plantbased", "plant milk" and "beyond meat"), we found no tweets, and 31 tweets possess more than one keyword; see Table 1.

Scraping the twitter

For this study, data was collected from Twitter between January 1st, 2010, and May 19th, 2022, using Python *snscrape* module (*JustAnotherArchivist, 2022*). The search was performed using the following keywords: "*vegan*", "*vegetarian*", "*veggie*", "*veganism*", "*plant-based*", "*cruelty-free*", "*plant milk*", "*tofu*". We refer to a keyword as a word that is present either in the form of a hashtag or anywhere in the body of a tweet. As we have seen in the Kaggle dataset, out of 1.6 million tweets, the tweets containing at least one of the keywords amounted to 1,876, which means a 0.0011725 content ratio. Thus, assuming 500 million tweets daily, we would have approximately 586,250 of our target tweets. Therefore, for the sake of getting the data fast, we limited the collection of data by 100 tweets for every keyword monthly along the 136 months and 19 days period collecting 113,677 tweets in total.[1]

[1]The data collection Python script is available at https://github.com/iskander-akhmetov/vegan_tweets_SA.

Among the collected tweets, 87% were in English, 3% in Japanese, and 2% in Spanish. We have also labeled tweets in other languages with less than 1000 tweets as "other"; additionally, Twitter has a category "und" for undefined languages; see Fig. 3. Besides English, there are numerous Japanese tweets, which signal that the vegan topic is trendy in Japan.

From Table 2, we see that 70% of the tweets do not have a like, reply, retweet, or quote and are entirely unnoticed. As for the most popular tweets, they comprise the top 10% by replies/retweets or having at least one reply or retweet, the top 30% by having at least one like, and the top 1% by having at least one quote.

Taking into account that the majority of the tweets have no likes, we have decided to analyze only the tweets that have at least one like. Thus, we have 36,959 tweets to explore; see Fig. 4 for the distribution of the liked tweets by the query.

After filtering out the tweets with no likes, we observed in the Fig. 5 that there had been a steady rise in the number of tweets on the "*vegan*" topic since 2012 from less than 1,000 tweets per year to more than 5,000 tweets in 2021. This can be explained by the fact, that

Table 1 Kaggle Sentiment140 dataset vegan keyword content.

Keyword	Tweet count
vegan	364
plantbased	0
healthy	955
vegetarian	191
veggie	391
veganism	5
cruelty free	1
plant milk	0
beyond meat	0

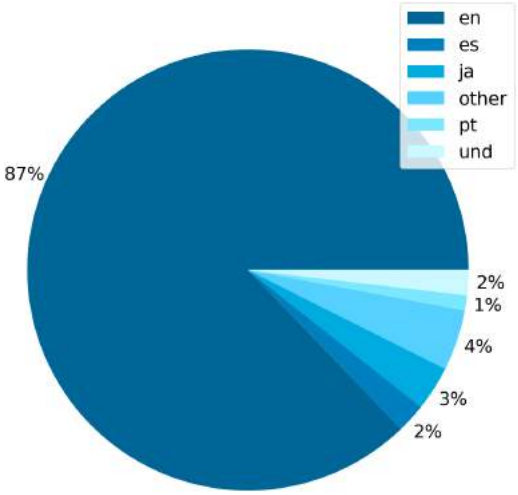


Figure 3 Scraped tweets share by language.
Full-size DOI: 10.7717/peerjcs.1149/fig-3

starting from 2011, a large amount of research has begun to appear on the health benefits of a plant-based diet. So, we suppose that the main reason for the rise of the trend in health care, besides the support of animal rights and environmental protection.

Moreover, there is a pattern of having relatively more tweets on the "*vegan*" subject during the spring months, which plummet in the middle of the summer and start rising in autumn again; see Fig. 6. The pattern can be explained by the fact that in the winter and spring months, people generally have a vitamin deficiency and start focusing on their health more.

Data description

The collected data set consists of 113,677 tweets. In the data set, 2,682 tweets correspond to health topics, containing healthy, wholesome, and full-blooded keywords. Also, there are three levels of vegan topic detailing, namely "vegan-1", "vegan-4", and "vegan-7". The level of "vegan-1" corresponds to just one keyword, namely "vegan", and there are 32653 tweets of it. The level of "vegan-4" is about "vegan, vegetarian, veggie, veganism,"

Table 2 A number of replies, retweets, likes, and quotes in the scraped tweets.

	Replies	Retweets	Likes	Quotes
count		**113,677**		
mean	0.33	0.53	3.14	0.05
std	3.92	23.02	344.35	4.20
min	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00
70%	0.00	0.00	1.00	0.00
80%	0.00	0.00	1.00	0.00
90%	1.00	1.00	3.00	0.00
95%	1.00	2.00	5.00	0.00
97%	2.00	3.00	9.00	0.00
99%	3.00	7.00	28.00	1.00
max	1,119.00	7,492.00	115,499.00	1,398.00

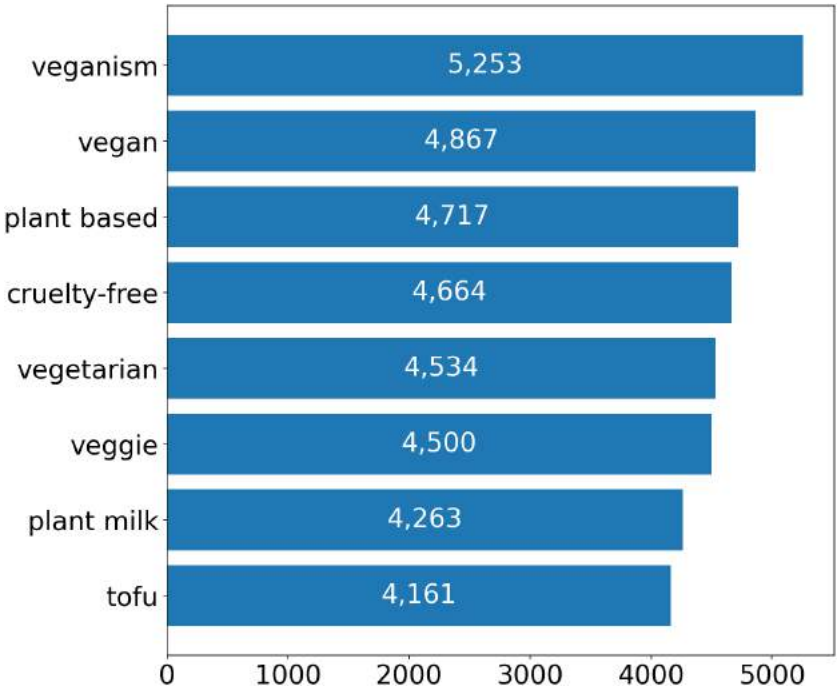


Figure 4 A number of tweets by the query.

Full-size DOI: 10.7717/peerjcs.1149/fig-4

and there are 53,201 tweets. The level of "vegan-7" is with the following keywords: "vegan, vegetarian, veggie, veganism, plant-based, plant-based, plant based" and there are 65,213 tweets, and 2,075 tweets include keywords about both veganism and health. Also, there are corresponding intersections between the health topic and the levels of the vegan topics: 1,359, 1,791, 2,252 tweets. The methodology of data collecting assumes that tweets are distributed equally in terms of years, months, and weekdays; however, the distribution by hour expects that users wrote the most tweets in the evening.[2]

[2]The data description Python script is available at https://github.com/AlexPak/VeganTwitterEDA/blob/main/Visualization.ipynb.

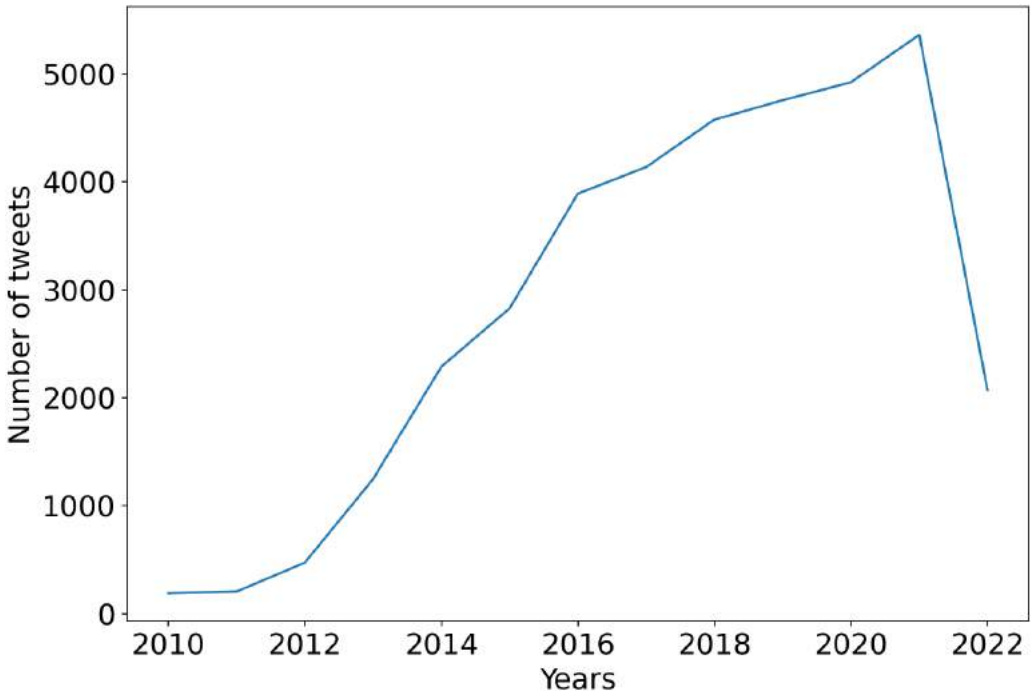


Figure 5 **Tweets having at least one like, by year from the 2010 to 2022 period.**
Full-size DOI: 10.7717/peerjcs.1149/fig-5

A visual representation of the keyword distribution can be seen in Fig. 7 from which it becomes clear which topics and discussions prevail in the presented data set. For example, the following topics: diets, recipes, animal welfare, etc.

We presented the cross-correlations between keywords in Fig. 8. You can see an anti-correlation between the keyword "plant milk" and the other keywords from the image. From the additional analysis, the cause of such phenomena is the insufficiency of plant-milk tweets against other subtopics, which leads to negative correlation values. Also, we should mention that there has been an increasing trend of mentions on Twitter "plant-milk" from 2010 to the present. Also, there is another systematic distortion in the cross-correlation diagram, particularly the high correlation value between vegan and veganism because one of them is a sub-string of another. All other features of the dataset are pretty much inter-correlated.

It is interesting to show the analysis of the relationship between health and vegan topics in a time-based manner. In Fig. 9, there is Spearman's rank correlation between the number of tweets dedicated to these topics.

There is a peak correlation with all topics in 2012. We should notice that the correlation with the topic of cruelty-free arises to its maximum with a one-year lag. The lowest maximum owns the keyword "vegan." The highest maximum owns the keyword "tofu." The next explicit historical maximum was in 2015; the highest value belongs to the keyword "vegetarian," and the lowest is the same as in 2012. In 2018 there is another peak with a different leader and outsider, namely the highest value owned to "tofu" and the lowest

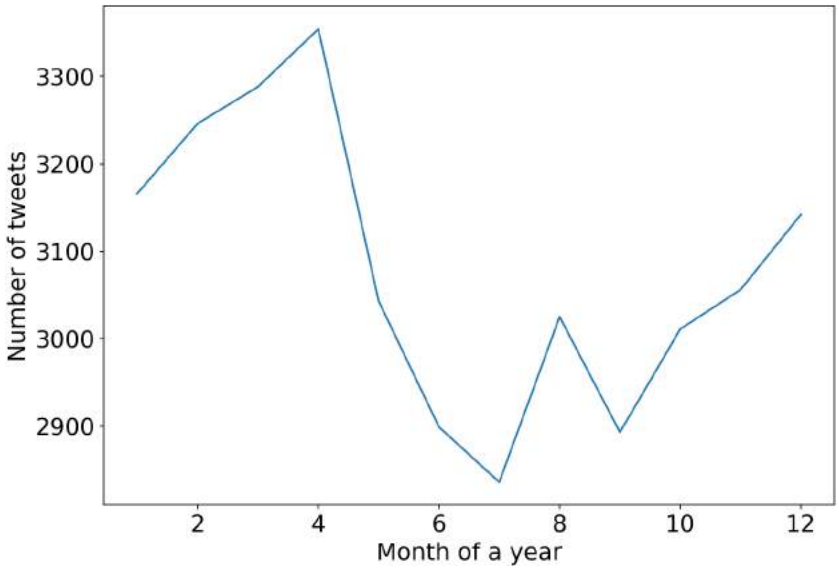


Figure 6 **Tweets having at least one like, by month of a year for the 2010 to 2022 period.**
Full-size DOI: 10.7717/peerjcs.1149/fig-6

one to a couple of keywords "plant-based, veggie." And the lowest peak all the time was in 2020; the leaders of the year are the keywords "plant-based, veganism", and the outsider is the keyword "tofu".

Another presented metric of discourse analysis of these topics is mutual information. From Fig. 9 we can see that the peak of simultaneous mentions of keywords from both topics is in 2018. Also, Fig. 9 shows an ascending trend from 2010 to 2018.

We strongly believe that such findings can give a clue to understanding the development of the social discourse on a couple of interrelated topics, "vegan" and "health."

Pre-processing

Twitter is an excellent source for public opinion mining. Yet, we must deal with its informal language, regular misspelling, wrong word order, etc. Therefore, Twitter data cannot be analyzed in its initial raw form. Twitter users type shortened words to maintain the word length limitation (*e.g.*, 140 characters maximum) in a tweet and tend to use digitally mediated language, so-called Slanguage (*Turney, 2002*), that incorporates slang, abbreviations that are natural in relaxed and unofficial context. In other words, tweets are required to be cleaned before the analysis processes. This process involves five steps: *Cleaning, Case Folding, Tokenization, Slang Lookup, and Stopword Removal.*

For better understanding, suppose we have a sample tweet in our data—"*I'm glad he responded to you. Too bad he thinks veganism is some kind of perfection or is difficult. Seriously. Difficult? C'mon:)*". The pre-processing steps are explained below and demonstrated in Fig. 10.

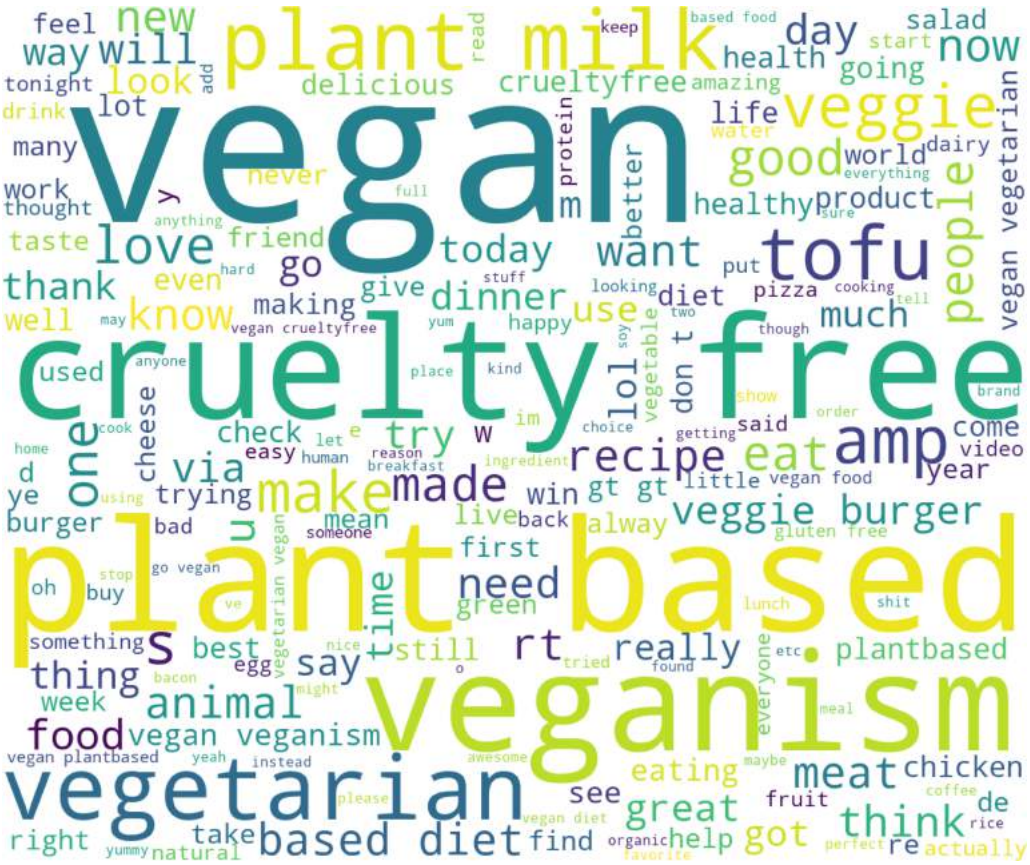


Figure 7 **Word cloud of the presented dataset with tweets in vegan context.**
Full-size DOI: 10.7717/peerjcs.1149/fig-7

Cleaning

There are a few specific types of unwanted textual parts that a standard NLTK tokenizer misses: mentions, emoticons, and URLs. At this step, we remove unnecessary text associated with them. The removed parts were: emoticons and punctuation marks.

Case folding

The primary purpose of this step is to convert the tweet to lowercase, thus minimizing the vocabulary size.

Tokenization

The tokenization step was done using NLTK tokenizer (*Bird & Klein, 2009*). We split all the words separated by a white space and remove all the punctuations and numbers. The tokenization process is based on regular expressions (regexp). Resultant tokens can be seen in Fig. 10.

Slang lookup

At this stage, we have the primary goal to remove slang words from the dataset. The term *slang* is a novel word that is not standard, and it is a dynamic, innovative, and ever-changing

phenomenon. During text analysis, slang words bring unsuitable implications. We apply the synonym method at this step to replace slang with its standard equivalent word (*Utama, 2019*). We do this by comparing the words in the dataset to the slang word list and synonyms and finding matches. In our example tweet, the slang words *C'mon, I'm* were modified.

Stopword removal

During the pre-processing stage, stop-word elimination is a crucial step. Not all of the words obtained at the previous stage are significant. Certain words have no sense, and they refer to stop words (articles, conjunctions, some adverbs, *etc.*). These terms should be removed from the dataset to speed up the processing time (*Utama, 2019*). Returning to our example, we removed common stop words, like *I, am, he, to, you, too, is, of, or, on*.

We examined the general structure of a tweet in the preceding explanation and how to pre-process its content. As a result, we have preprocessed tweets in the form of clean tokens that can be further passed to perform mutual information feature selection.

Sentiment detection

Computing term probabilities

Computation of $P(t)$ (the probability of observing the term t) and $P(t_1 \wedge t_2)$(the probability of co-occurring of the terms $t1$ and $t2$) is defined in Eqs. (1) and (2) (*Turney & Littman, 2003a*; *Turney & Littman, 2003b*; *Turney, 2002*).

$$P(t) = \frac{DF(t)}{|D|}, \tag{1}$$

where D is the number of documents (or tweets in our case) in our dataset, and DF is Document Frequency or number of documents where the term occurs (*Bonzanini, 2016*).

$$P(t_1 \wedge t_2) = \frac{DF(t_1 \wedge t_2)}{|D|} \tag{2}$$

Mutual information feature selection

By deleting uninformative characteristics, feature selection tries to minimize classification calculation time. Mutual information is a machine learning selection approach demonstrating how important a feature is in making an accurate prediction (*Bonzanini, 2016*).

We use mutual information to determine the strength of a semantic association between two words. We compute the PMI of each pair of terms and then calculate the semantic orientation given two vocabularies for positive and negative terms. If some word is commonly associated with concepts in the positive lexicon, it has a positive semantic orientation. A schematic representation of the method can be seen in Fig. 11.

To begin, we define a word's semantic orientation (SO) as its positive and negative connections difference (*Turney & Littman, 2003a*; *Turney & Littman, 2003b*; *Turney, 2002*). In practice, we're looking for a way to determine "how near" a word is to phrases like good and awful. Point-wise mutual information (*PMI*) is the preferred measure of "closeness," which is defined in Eq. (3) (*Church & Hanks, 2002*; *Turney & Littman, 2003a*;

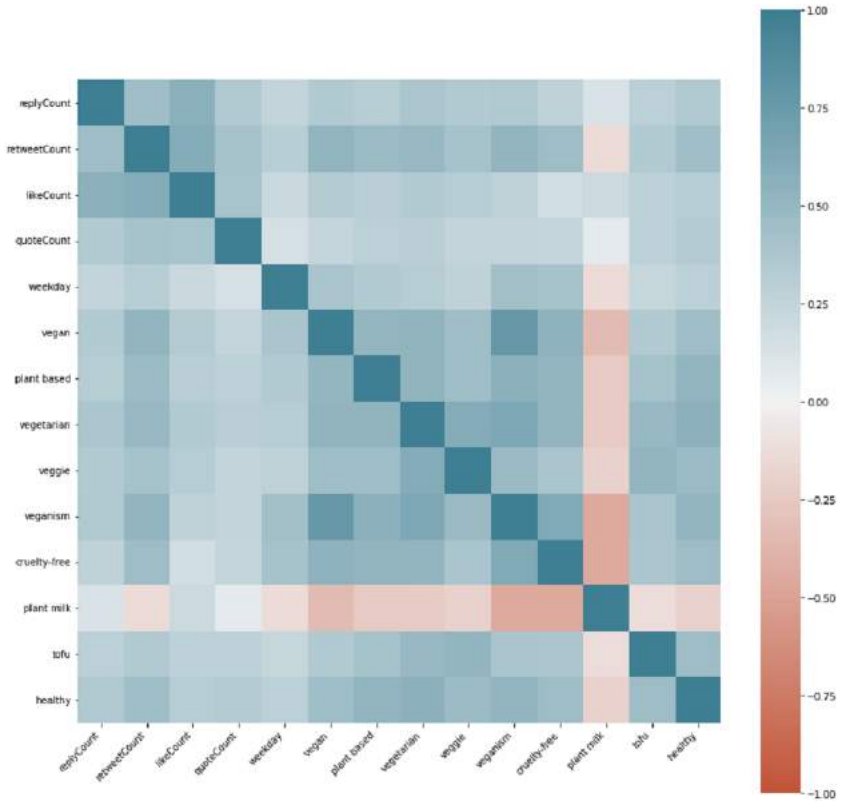


Figure 8 **Cross-correlation between the features of the data-set.**
Full-size DOI: 10.7717/peerjcs.1149/fig-8

Turney & Littman, 2003b; *Turney, 2002*).

$$PMI(t_1, t_2) = log\left(\frac{P(t_1 \wedge t_2)}{P(t_1) \times P(t_2)}\right), \tag{3}$$

where t_1 and t_2 are terms and P designates probability function.

The semantic orientation (SO) of a term t is defined in Eq. (4).

$$SO(t) = \sum_{t' \in V^+} PMI\left(t, t'\right) - \sum_{t' \in V^-} PMI\left(t, t'\right), \tag{4}$$

where V^+ and V^- are vocabularies of positive and negative words respectively.

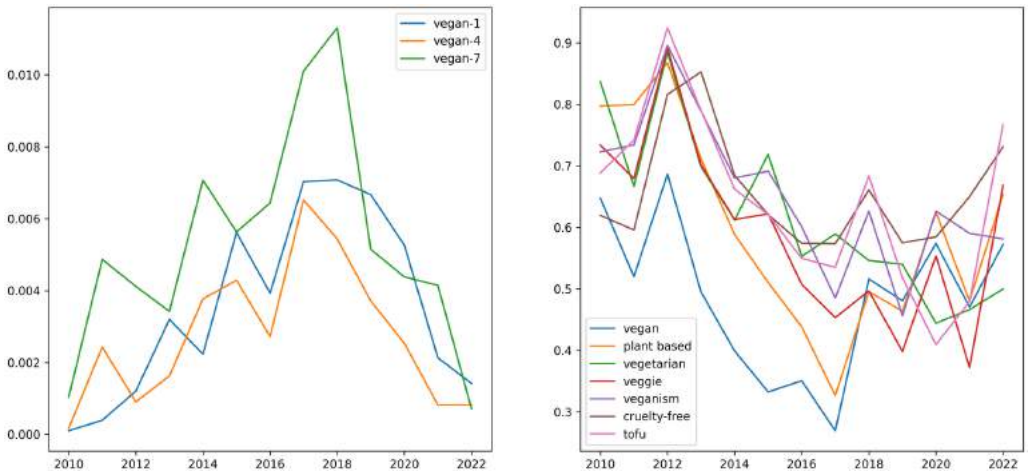


Figure 9 **The time-based evaluation of mutual information on two topics, health and vegan (left). The time-based correlation between the topic of health and vegan topic keywords (right).**
Full-size DOI: 10.7717/peerjcs.1149/fig-9

Semantic orientations of some context-specific words from our collected dataset are presented in Fig. 12. As we can see from the figure, many words specifying animal products have negative semantic orientations in the vegan context. In contrast, the words associated with plant-based diets mostly have a positive semantic orientation. It is of interest to note that the word "eggs" has a far more negative score than the word "beef", for example (−1.716 and −0.226, respectively). Note that these terms may have different scores when discussed in non-vegan contexts. At this step, we can update the general lists of positive and negative words we initially had with context-specific words from our dataset that got very high or very low scores, respectively.

The corresponding Algorithm 1 along with the pseudocode can be seen below. Python code and all supplemented data are also publicly available.[3]

[3]Sentiment analysis Python code and all supplemented data is available at https://github.com/elvinast/Sentiment-Analysis-Vegan-Tweets

Algorithm 1: Mutual information for feature selection

```
Data: tweets in csv file
Result: obtained PMI for semantic orientation calculation;
df_clean["tokens"] = tokenized tweets by words;
remove stopwords and change slang words;
Function build_frequencies(tweets):
    word_count = calculated count of each word;
    word_pairs = calculated word pairs count;
    return (word_count, word_pairs);
Function build_probabilities(word_count, word_pairs, tweets_count):
    word_probability = dictionary that contains calculated probability of each word;
    word_pair_probability = dictionary that contains calculated probability of each
    pair;
    return (word_probability, word_pair_probability);
Function build_pmi(tweets):
    word_count, word_pairs = build_frequencies(tweets)
    word_probability, word_pair_probability =
    build_probabilities(word_count, word_pairs, tweets_count) pmi ← emptydictionary
    for term1 in word_probability do
        for term2 in word_pairs[term1] do
            denom = word_probability[term1] * word_probability[term2]
            pmi[term1][term2] calculated by formula 1
        end
    end
    return pmi, word_count;
// Load set of positive and negative words
Function get_semantic_orientation(word):
    pmi, word_count = build_pmi(df_clean["tokens"])
    if word in word_count then
        return value calculated by Formula 2
    end
```

Classifying tweets

Furthermore, we can find a tweet sentiment by finding a semantic orientation of cleaned tokens from each tweet in a training dataset. The sentiment score of a tweet is calculated as the sum of the sentiment scores of each term present in the tweet and in the lexicon (Eq. (5)) (*Bindal & Chatterjee, 2016*):

$$TweetScore(T) = \sum_{t \in T} SO(t), \tag{5}$$

where T is a tweet, t - specific term in a tweet, $SO(t)$ is a term score (semantic orientation of t, Eq. (4)). Note that we do not consider the scores for out-of-lexicon terms. The classification threshold for neutral tweets was set from -0.1 to 0.1. The tweets with negative scores were assigned a negative sentiment, whereas positively scored tweets got

a positive sentiment. The scores close to 0 represent neutral tweets. It happens when the PMIs for positive and negative balance out.

The accuracy of the proposed method was 0.62 on a training dataset (labeled tweets from Kaggle). This indicates the goodness of the sentiment lexicon, considering that neither this dataset was manually annotated by humans nor the classifier used for labeling was trained on vegan-related language. We will continue discussing the accuracy in the Results section when we utilize a dataset with vegan-related tweets labeled by humans.

Emotion detection

Emotion recognition is frequently considered a subtask of polarity detection (*Cambria, 2016*), and it became the next step naturally after the basic sentiment task (*Ashraf et al., 2022*). A user puts into the tweet not just a positive or negative meaning but often a range of emotions. Other users can easily recognize and experience these emotions in texts. Emotion investigation can help us better understand opinions rooted in tweets. Detection of emotions behind the tweets was done using *text2emotion 0.0.5* (*Python Software Foundation, 2020*). Five different emotion categories were used, such as *Happy, Angry, Sad, Surprise and Fear.* Text2emotions returns the dictionary with scores for each emotion category.

We could assume that the tweet belongs to a particular emotion if the score for that emotion is high. However, we don't want to miss emotions of average and weak power. So we use a fuzzy approach (*Zadeh, 1965*) to manipulate emotions expressiveness.

A class of objects with a continuum of grades of membership is referred to as a fuzzy set (*Zadeh, 1965*; *Zadeh, 1975*). Several studies used a fuzzy approach to represent the strength or type of emotions and differentiate between different emotional labels (*Matiko, Beeby & Tudor, 2014*; *Karyotis et al., 2016*; *Qamar & Ahmad, 2015*).

The primary motivation for using fuzzification in our study is that it allows us to judge emotions in a human-consistent style. Since emotions don't have clearly defined boundaries, fuzzy sets are very useful for dealing with these concepts. By partitioning the range of feasible emotions corresponding to linguistic tags, we get more intuitive and human-consistent output (*Shamoi et al., 2016*). Likewise, fuzzy sets can easily depict the gradual change from one emotion label to another can (*Zadeh, 1975*). A language value is more similar to human cognitive processes than a number, despite being less precise (*Shamoi & Inoue, 2012*).

We represent *Emotion power* as an ordered list of terms of the fuzzy (linguistic) variable $X = \text{“}Emotion\ power\text{”}$ by means of primary linguistic terms $L = \{Weak, Medium, Strong\}$, specifying its level. Figure 13 below depicts the fuzzy sets for the *Emotion power*. For the sake of simplicity, we use triangular membership functions (Eq. (6)) representing the linguistic labels expressing the fuzzy variable *Emotion power* for each of the emotion categories. It is good enough to catch the ambiguity of the linguistic assessments. Such fuzzy partition was made based on subjective perception —we just ran a few experiments and adjusted the fuzzy sets. In the future, we can improve the partition by conducting a survey based on human emotion categorization (*Shamoi, Inoue & Kawanaka, 2014*).

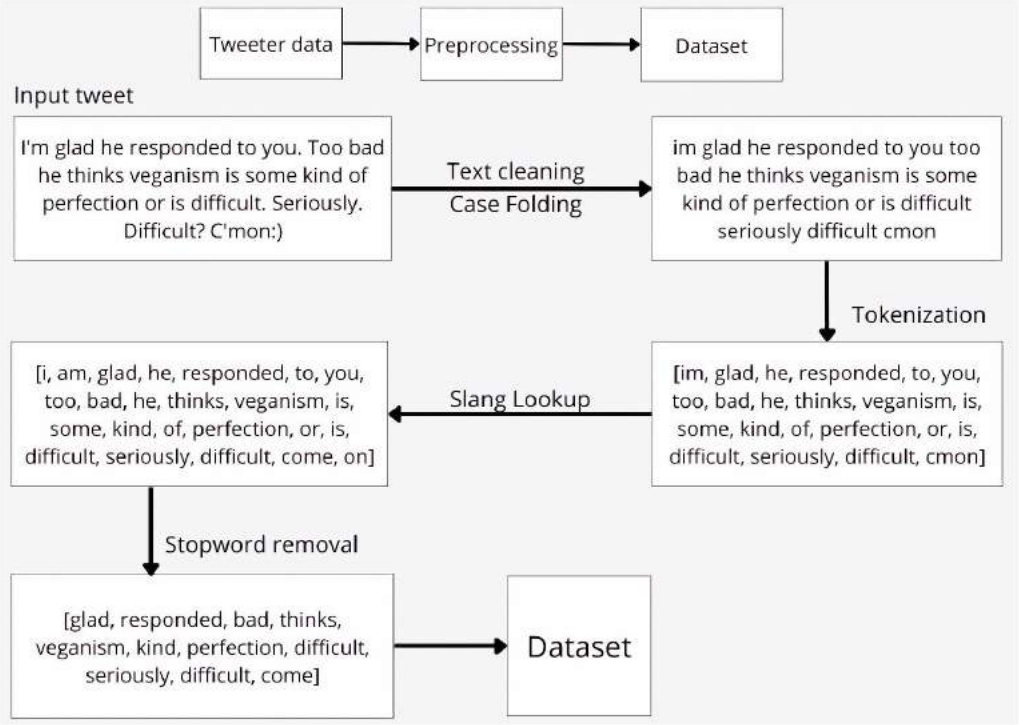


Figure 10 **Sample tweet.** The pre-processing steps.
Full-size DOI: 10.7717/peerjcs.1149/fig-10

$$\mu_a(x) = \begin{cases} 0, & \text{if } x \leq a. \\ \dfrac{x-a}{b-a}, & \text{if } a \leq x \leq b\,. \\ \dfrac{c-x}{c-b}, & \text{if } b \leq x \leq c. \\ 0, & \text{if } x \geq c, \end{cases} \tag{6}$$

where a, b, c represent the x coordinates of the three vertices of $\mu_a(x)$ in a fuzzy set A. Although we have five types of emotions, not all of them are present in a tweet. We take into account only those that obtained a non-zero score. For example, some specific tweet may contain *weak sadness* and *strong surprise* emotions only.

In the case of composite emotion impressions, formed from atomic ones with the help of various connectives (and, or) (*Shamoi & Inoue, 2012*) and hedges (*e.g.*, not, very, more-or-less), general formulas from fuzzy theory, like α-cut, reinforcing and weakening modifiers, set operations (*Shamoi, Inoue & Kawanaka, 2020*) are used. α-cut is a crisp set that includes all the members of the given fuzzy subset whose values are not less than α for $0 < \alpha \leq 1$. The α-cut of A is defined as *Zadeh (1965)*:

$$[A]_\alpha = \{x \in X | A(x) \geqslant \alpha\}. \tag{7}$$

These basic formulas allow us to retrieve some interesting tweets and judge emotions in human-consistent style, like *"extremely happy and surprised"*, *"very very angry but not sad"*, etc.

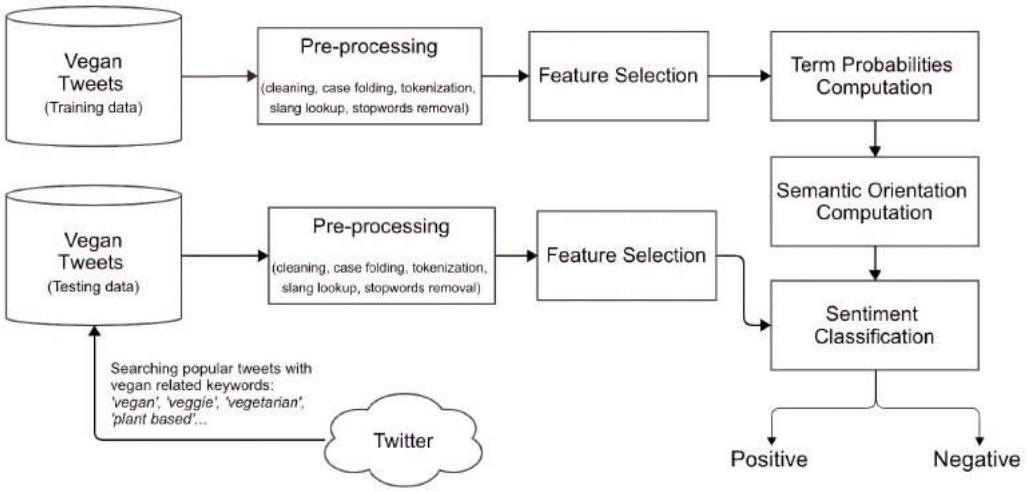


Figure 11 **Schematic representation of the model.**
Full-size DOI: 10.7717/peerjcs.1149/fig-11

As we see, emotion analysis can help to detect more complex feelings embedded in a tweet. For example, '*surprise*' is an emotion that can express both positive and negative sentiments (*Sailunaz & Alhajj, 2019*). The negative form of *surprise* is when you eat well, yet the tests reveal nutrient deficiency. On the other hand, the positive emotion of *surprise* is when you are surprised by an extensive choice of vegan options in a restaurant.

Let us take, for example, a tweet *"Sorry Vegans, steak taste good."*. Its sentiment score is *Neutral* (−0.043) due to its balance of positive and negative words. At the same time, emotion analysis results are *'Happy': 'Medium', 'Sad': 'Weak', 'Fear': 'Weak'*. So, detecting emotions can be particularly useful for balanced tweets that standard SA can neutralize.

Python code and all supplemented data for emotion detection is publicly available.[4]

[4]Python code and all supplemented data is available at https://github.com/elvinast/Sentiment-Analysis-Vegan-Tweets.

Dataset annotation on an 11-point scale

We initiated the annotation to evaluate the method because there are no available vegan-specific datasets labeled by humans or by a classifier trained in the vegan context. The trial dataset consisted of 100 tweets and was created using random sampling from our vegan-related unlabeled dataset described before.

Five humans performed the annotation process. There were three vegans and two omnivores among them. Data annotation should be performed by someone with a thorough understanding of the topic, so local vegans with 15, 5, and 2 years of vegan experience were asked to contribute. Vegans were given double weight since they are more familiar with the lexicon of the context. The human raters used separate Google sheets they got online to label the sentiment polarity of the tweets. A screenshot of the form (Google Sheets) is provided in Fig. 14.

Each tweet in a dataset was assigned a category by all five annotators. A test dataset was annotated using an 11-point scale annotation method described in *Ghosh et al. (2015)*. Each annotator was instructed to assign a score to an opinion in a tweet based on a

perceived meaning between −5 (representing strong displeasure) and +5 (for strongly positive tweets). Details can be seen in Fig. 14.

We used tactics given in *Ghosh et al. (2015)*—If 60% or more of annotator labels are assessed to be outliers, then the annotator judgments are dismissed from the task. We use this formula Eq. (8) to identify if a judgement $A_{i,j}$ is an outlier: (*Ghosh et al., 2015*):

$$|A_{i,j} - avg(A_{i',j})| > std_t(t_j), \tag{8}$$

where $std_t(t_j)$ is the standard deviation of all scores given for a tweet t_j.

After the weighted average calculation, we revealed one annotator whose proportion of outliers was 71%, so his annotations were excluded from the analysis, and the weighted average for each tweet was computed again.

Eventually, based on weighted average scoring, each tweet in a trial dataset obtained either a positive, neutral, or negative score (see Table 3). The annotation procedure resulted in the following sentiment distribution: positive sentiments (57%), neutral sentiments (0%), and negative sentiments (43%), with classification thresholds set at −0.1 and +0.1.

EXPERIMENTAL RESULTS AND DISCUSSION

Sentiment detection results

Some tweets lead to discussions involving a lot of users. Popular tweets are tweets with relatively high responses in the form of retweets, comments, and likes (*Zubiaga et al., 2014*). For sentiment analysis, we selected popular tweets from our vegan dataset for each year from 2010 to 2022. By popular tweets, we mean those tweets that were retweeted at least once, indicating that the opinion expressed in them was supported by other users.

Obtained sentiment analysis results are presented in Fig. 15 and Table 4. Results indicate that positive sentiment in popular tweets consistently dominates over the years. It can also be seen that all sentiments started to grow, particularly during those years when the associated studies on the health benefits of a plant-based diet appeared. As for 2022, we see a fall in the number of popular tweets, which is associated with the fact that our dataset included only 4.5 months for 2022. However, it is evident that in 2022 the proportion of tweets with negative sentiment relatively decreased compared to previous years and is nearly equal to the proportion of neutral tweets. This could be interpreted in several ways, the most common of which is the impact of COVID-19 and global warming. Plant-based foods are becoming more common in post-pandemic diets. The investigation of the 'Popular Tweets' column of Table 4 reveals that levels of interest in plant-based diets have increased from 2010 to 2022. Note that both Fig. 15 and Table 4 focus only on tweets that have been retweeted at least once (RT $\geq$ 1).

Table 5 shows examples of some positive, neutral, and negative tweets from our dataset along with their computed scores using Eq. (5).

Emotion detection results

Sentiment analysis helped us to identify that positive sentiments were prevailing. However, what are the leading emotions behind this growing interest, and how were they changing throughout the period?

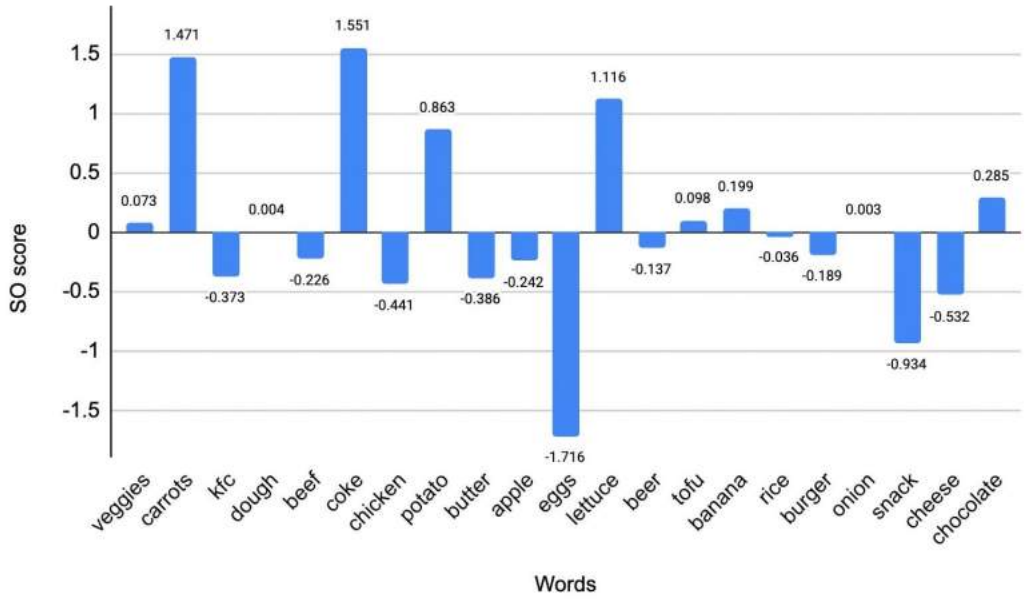


Figure 12 **Semantic orientations of some context-specific words.** This illustration can give us an idea of the main topics/ words when veganism is discussed in a positive or negative context. As can be seen, there is plenty of healthy food words which are scored as positive in vegan tweets.
Full-size DOI: 10.7717/peerjcs.1149/fig-12

The graphs in Fig. 16 show the comparisons of emotions across years that indicate some interesting facts. The release of a lot of popular documentary movies revealing violence towards animals in agriculture and the health benefits of plant-based diets can explain the spike in all emotions in 2018 (What the Health, 2017 (7.2/10 IMDb); Game Changers, 2018 (7.8/10 IMDb); Dominion, 2018 (9/10 IMDb); Eating You Alive, 2018 (7.9/10 IMDb); 73 cows, 2018 (7.8/10 IMDb); Let Us Be Heroes, 2018 (8.7/10 IMDb)). Moreover, in 2018, UN IPCC reported that climate change catastrophe would occur in the upcoming 12 years (*Masson-Delmotte et al., 2019*).

The growth of *Happy* emotion, which we can clearly observe from 2018, is possibly connected with the fact that more and more vegan options, products, and places are starting to open, and people feel joyful about their health, lifestyle, and food. The '*surprise*' emotion noticeably increased from 2018 as well, which can be connected to the fact that documentaries about the benefits of the vegan diet and the truth about animal abuse premiered this year (What the Health, 2017 (7.2/10 IMDb); Game Changers, 2018 (7.8/10 IMDb); Dominion, 2018 (9/10 IMDb); Eating You Alive, 2018 (7.9/10 IMDb); 73 cows, 2018 (7.8/10 IMDb); Let Us Be Heroes, 2018(8.7/10 IMDb)).[5] This fact can also be considered as the reason for the rise of the *Sad* emotion because most people feel heartbroken, frustrated, and sad about realizing the heavy burden that animals inhabit for just one human's meal. Furthermore, on the graph, we can recognize that emotions of *Fear* were mostly medium and strong throughout the period, which contrasts sharply with other types of emotions. This can be explained by the fact that many people tend to believe that a vegetarian or vegan diet can lead to nutritional deficiency (*Paslakis et al., 2020*).

[5] Information is taken from official IMDB site https://www.imdb.com/.

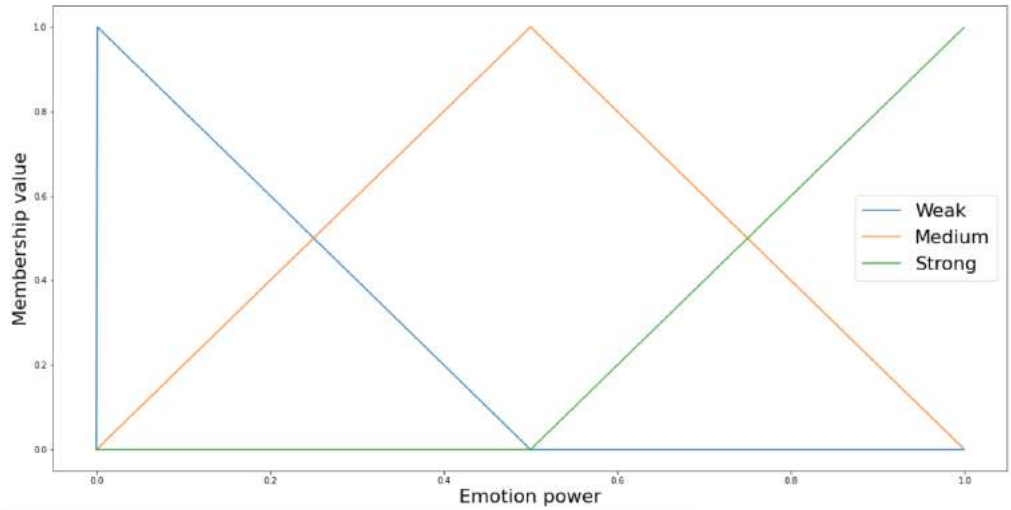


Figure 13 **Fuzzification of emotions levels.**
Full-size DOI: 10.7717/peerjcs.1149/fig-13

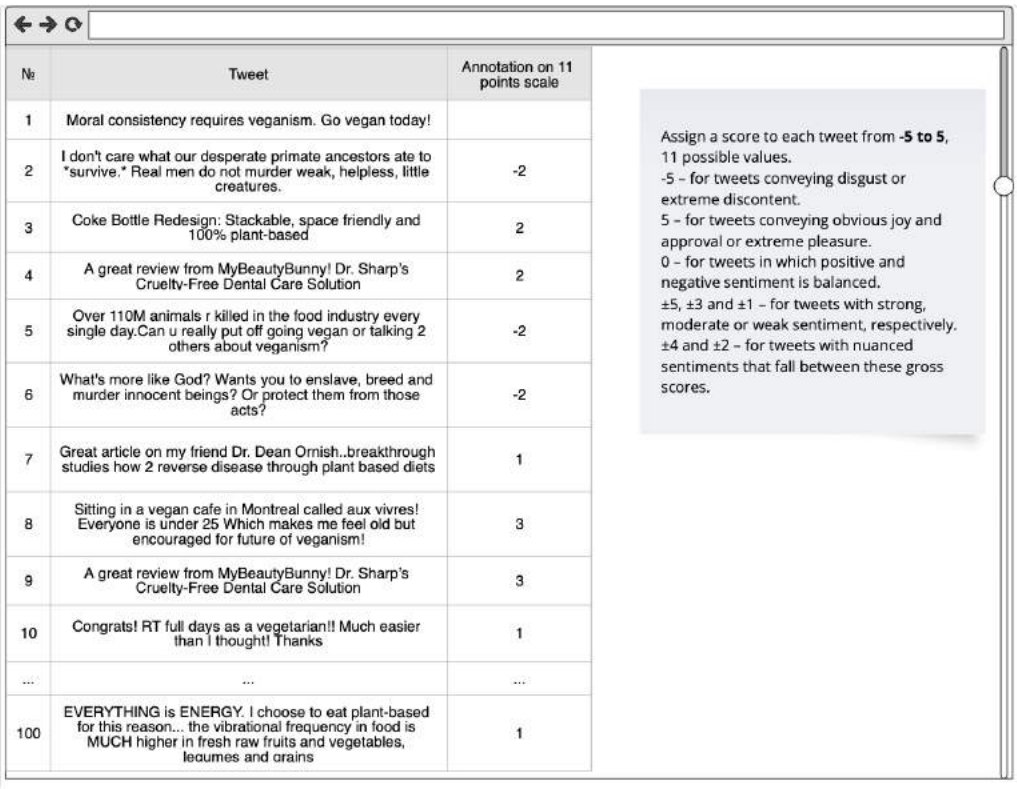
Table 3 **Examples of sentiment results for some tweets based on manual annotation.** Evaluations provided by Annotator A3 were excluded from the weighted average calculation as the number of outliers in its results exceeded 60%. Annotators with asterisks were given a double weight because they are vegans and are good at a context-specific lexicon.

#	Tweet	A1* $w=2$	A2 $w=1$	A3 (Excluded)	A4* $w=2$	A5* $w=2$	Weighted AVG	Annotation result	Sentiment
1	Moral consis...	1	3	5	1	4	2.14	1	Positive
2	I don't care ...	−3	1	3	−2	−4	−2.43	−1	Negative
3	Coke Bottle ...	3	4	5	2	3	2.86	1	Positive
4	A great revie...	3	4	0	2	3	2.86	1	Positive
5	Over 110M a...	−3	0	5	−2	−3	−2.29	−1	Negative
6	What's more...	−5	0	5	−2	−2	−2.57	−1	Negative

Figure 17 demonstrates how individuals perceived veganism over time. We investigated that the public's interest in plant-based diets has been growing noticeably since 2010. In 2017-2022, compared to 2011-2016, it can be noticed that the number of *weak happy* and *weak angry* tweets doubled. At the same time, there were no significant changes in the amounts of *medium happy*, *angry* and *strong happy* tweets. But it should be noted that there were fewer *strong angry* tweets by 2017–2022. From this, it can be assumed that by 2017–2022, people's attitude to plant-based lifestyles was changing for the better.

Table 6 illustrates examples of emotions identified in tweets from our dataset.

We also performed a basic analysis based on the usage frequency of words in a vegan context across 12 years. The bar chart (see Fig. 18) illustrates the most frequently used terms (excluding keywords we used for our dataset). It can be clearly seen that when people discuss veganism, they usually don't just talk about food *(meat, products, recipe, organic,*



№	Tweet	Annotation on 11 points scale
1	Moral consistency requires veganism. Go vegan today!	
2	I don't care what our desperate primate ancestors ate to "survive." Real men do not murder weak, helpless, little creatures.	-2
3	Coke Bottle Redesign: Stackable, space friendly and 100% plant-based	2
4	A great review from MyBeautyBunny! Dr. Sharp's Cruelty-Free Dental Care Solution	2
5	Over 110M animals r killed in the food industry every single day.Can u really put off going vegan or talking 2 others about veganism?	-2
6	What's more like God? Wants you to enslave, breed and murder innocent beings? Or protect them from those acts?	-2
7	Great article on my friend Dr. Dean Ornish..breakthrough studies how 2 reverse disease through plant based diets	1
8	Sitting in a vegan cafe in Montreal called aux vivres! Everyone is under 25 Which makes me feel old but encouraged for future of veganism!	3
9	A great review from MyBeautyBunny! Dr. Sharp's Cruelty-Free Dental Care Solution	3
10	Congrats! RT full days as a vegetarian!! Much easier than I thought! Thanks	1
...	...	...
100	EVERYTHING is ENERGY. I choose to eat plant-based for this reason... the vibrational frequency in food is MUCH higher in fresh raw fruits and vegetables, legumes and grains	1

Assign a score to each tweet from **-5 to 5**, 11 possible values.
-5 – for tweets conveying disgust or extreme discontent.
5 – for tweets conveying obvious joy and approval or extreme pleasure.
0 – for tweets in which positive and negative sentiment is balanced.
±5, ±3 and ±1 – for tweets with strong, moderate or weak sentiment, respectively.
±4 and ±2 – for tweets with nuanced sentiments that fall between these gross scores.

Figure 14 Sample form (Google Sheets) with tweets for annotation that was provided separately to each annotator. Score descriptions and data annotation method were taken from *Ghosh et al. (2015)*. The form contains 100 random tweets from our unlabeled vegan dataset.
Full-size DOI: 10.7717/peerjcs.1149/fig-14

delicious), but also touch on the topics of *animals, health,* and *beauty* (perhaps this is related to the topic of testing products on animals).

Our findings support the results of the effects of sentiment on information diffusion, obtained by *Ferrara & Yang (2015)*. According to their findings, positive posts reach a wider audience, implying that people are more likely to retweet positive content, a phenomenon known as *positive bias*. Discussions with more than 200 tweets are considered active ones (*Ferrara & Yang, 2015*). We filtered out only active, extremely popular tweets having at least 200 retweets (RT $\geq$ 200). Next, we fit the data within a polynomial function as depicted in Fig. 19. It illustrates the number of retweets collected by the tweets as a function of the sentiment score expressed therein. As seen from the figure, among such viral tweets, we have primarily neutral and slightly positive tweets, but not extremely positive ones. Nevertheless, the average number of retweets is higher for negative tweets (0.706, 0.503, and 0.369 for negative, positive, and neutral tweets, respectively).

Accuracy evaluation

We use popular sentiment analysis models VADER (*Honnibal & Montani, 2017*) and spaCy (*Honnibal & Montani, 2017*) and the human-annotated dataset discussed earlier to evaluate the performance of our approach.

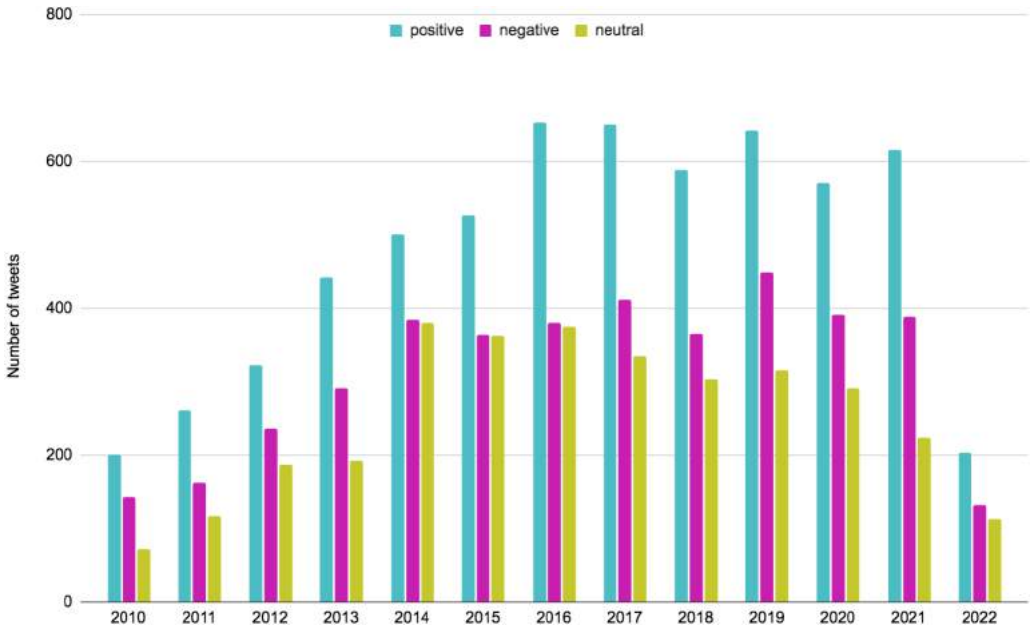


Figure 15 Sentiment analysis results for popular tweets (retweeted at least once, RT ≥ 1) from 2010 to 2022.

Full-size DOI: 10.7717/peerjcs.1149/fig-15

Table 4 Sentiment analysis results for popular tweets (retweeted at least once, RT ≥ 1) from 2010 to 2022.

Year	Popular tweets	Positive	Positive, %	Negative	Negative, %	Neutral	Neutral, %
2010	414	200	48.31	142	34.3	72	17.39
2011	542	262	48.34	163	30.07	117	21.59
2012	745	322	43.22	236	31.68	187	25.1
2013	927	443	47.79	291	31.39	193	20.82
2014	1,265	501	39.6	384	30.36	380	30.04
2015	1,254	527	42.03	364	29.03	363	28.95
2016	1,409	654	46.42	380	26.97	375	26.61
2017	1,397	651	46.6	412	29.49	334	23.91
2018	1,257	588	46.78	366	29.12	303	24.11
2019	1,407	643	45.7	449	31.91	315	22.39
2020	1,253	571	45.57	391	31.21	291	23.22
2021	1,227	615	50.12	389	31.7	223	18.17
2022	448	203	45.31	132	29.46	113	25.22

Table 5 Examples of some positive, neutral, and negative tweets from our dataset along with their computed scores.

Tweet	Score	Sentiment
"Sitting in a vegan cafe in Montreal called aux vivres! Everyone is under 25 Which makes me feel old but encouraged for future of veganism!"	1.65	Positive
"You can't convince anyone of the benefits of vegetarian/veganism when your diet consists of multiple fake meats. Just admit you want meat."	−0.15	Negative
"Stupid Questions (and clever answers) about veganism never gets old"	−0.27	Negative
"My boss wants to make me a vegan soy burger. Whoa slow down there."	−0.02	Neutral
"Enjoying the vegan treats at Café Petit Gteau"	0.72	Positive
"#Flexitarian diet: mostly plant-based with small portions of meat, fish and poultry."	0.02	Neutral
"my healthy goals are to lose 10 pounds, drink more water and use more plant-based foods."	0.9	Positive

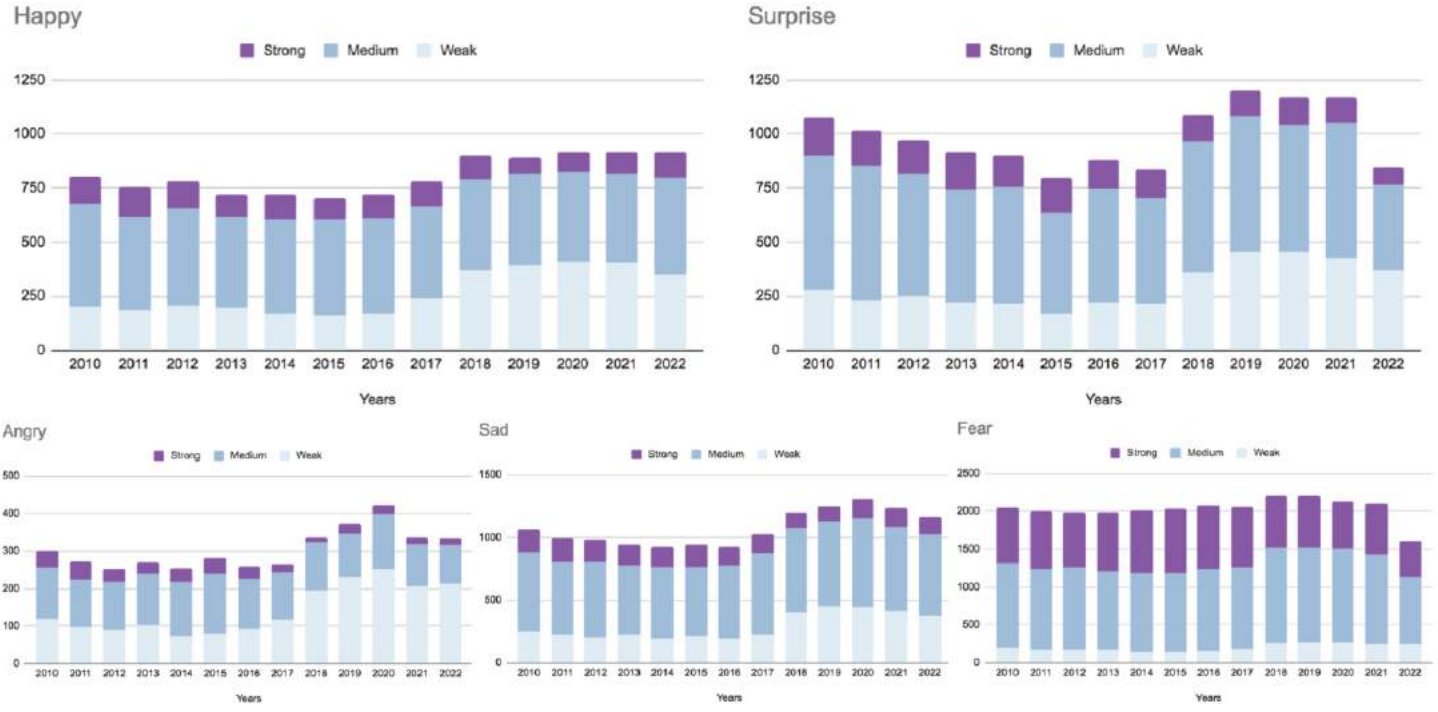


Figure 16 Prevailing emotions in tweets across years.

Full-size DOI: 10.7717/peerjcs.1149/fig-16

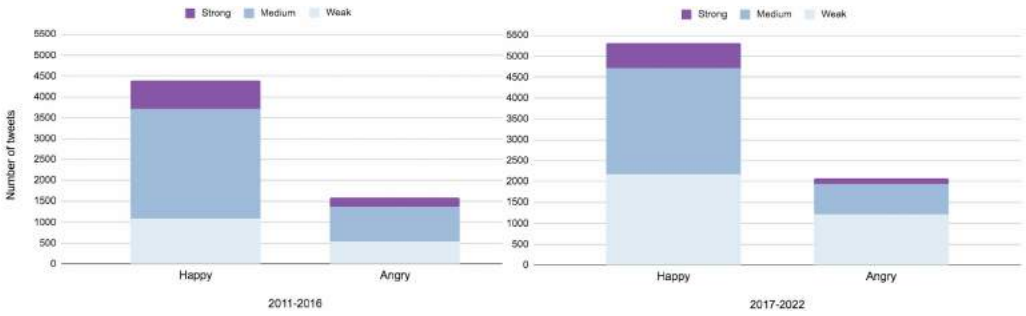


Figure 17 **Bar charts for happy and angry emotions for a 5-year subsequent period.**
Full-size DOI: 10.7717/peerjcs.1149/fig-17

Table 6 **Examples of fuzzy emotion detection in tweets from our dataset.**

Sample tweet	Crisp emotions	Fuzzified emotions
"Always puzzles me why so many plant based menu items reference non-plant based foods."	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.2, 'Sad': 0.6, 'Fear': 0.2}	{ 'Surprise': 'Weak', 'Sad': 'Medium', 'Fear': 'Weak'}
"handsome, musician, tattooed and vegetarian I'm already in love"	{'Happy': 1.0, 'Angry': 0.0, 'Surprise': 0.0, 'Sad': 0.0, 'Fear': 0.0}	{'Happy': 'Strong'}
"Meat is bad for you. Emissions from agriculture are projected to increase 80% by 2050. Being vegan makes a difference."	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, 'Sad': 0.67, 'Fear': 0.33}	{'Sad': 'Medium', 'Fear': 'Medium'}
"So if you don't want to shop cruelty free for the animals, do it for your own well being"	{'Happy': 0.2, 'Angry': 0.4, 'Surprise': 0.0, 'Sad': 0.4, 'Fear': 0.0}	{'Happy': 'Weak', 'Angry': 'Medium', 'Sad': 'Medium'}
"I became a vegetarian in college when I got very sick from eating meat. I really don't miss eating meat."	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.17, 'Sad': 0.33, 'Fear': 0.5}	{'Surprise': 'Weak', 'Sad': 'Medium', 'Fear': 'Medium'}
"I Thought u were a vegetarian"	{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 1.0, 'Sad': 0.0, 'Fear': 0.0}	{'Surprise': 'Strong',}

Table 7 presents the Mutual information scoring function against VADER and spaCy scorings along with the sample tweets from the labeled dataset.

Table 8 shows the three-class classification accuracy for each method. From Table 8 we can see that our classifier (0.77) performs better than VADER (0.69) and spaCy (0.67) for the selected context, but still not as well as individual human rater with special context knowledge (0.87). Unsurprisingly, our scoring method obtained much higher accuracy (0.77) than the accuracy on the training data set labeled automatically without accounting for the vegan-tweets-based lexicon. Although we know that VADER is sensitive to social media lexicon, its accuracy decreases since we use it in domain-specific vegan tweets. The dataset we used for training has narrow variability and context-specific vocabulary, so a PMI-based classifier performs better.

Next, let us look at how well different classifiers agree on the sentiments of the tweets. We calculated the intersections of each method's positive or negative proportions to compute the agreement. The *Agreement Score* represents the fraction of tweets both methods agree (have the same predicted polarity) (*Gonalves et al., 2013*). Table 9 shows the percentage of

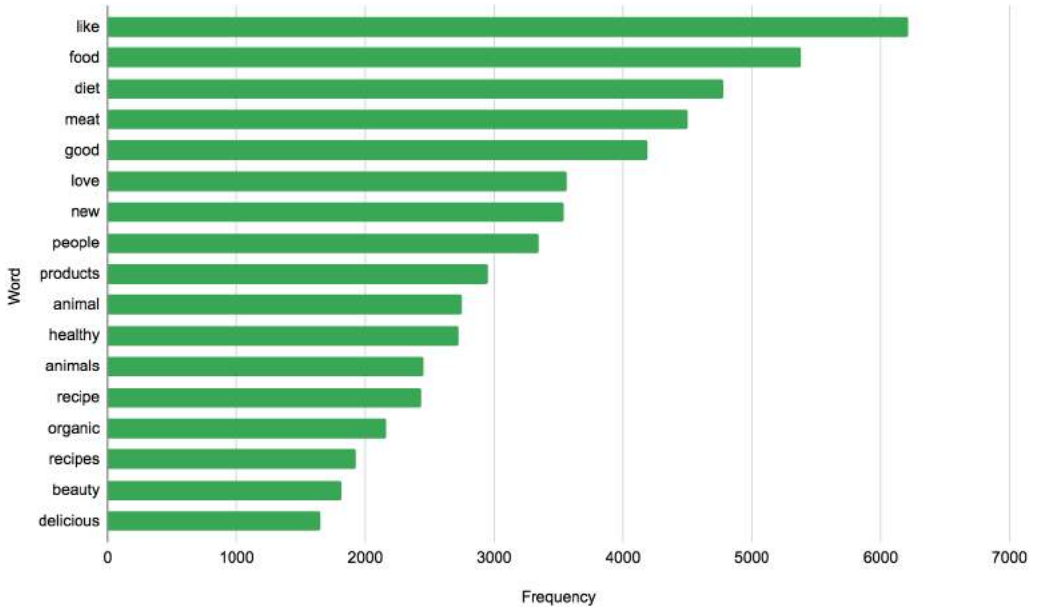


Figure 18 **Top frequent words in a dataset (excluding search keywords).**
Full-size DOI: 10.7717/peerjcs.1149/fig-18

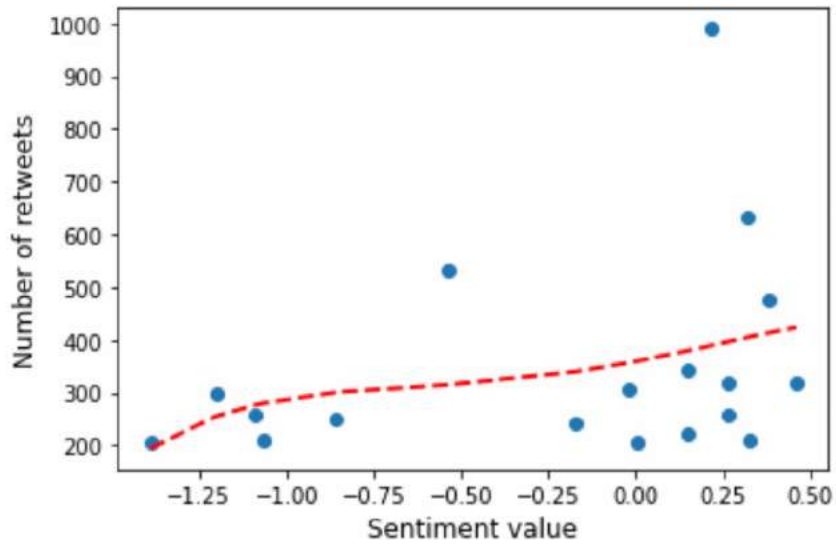


Figure 19 **Extremely popular tweets (RT ≥ 200) and their sentiment values.**
Full-size DOI: 10.7717/peerjcs.1149/fig-19

agreement between methods (phenomenon of methods agreement is discussed in *Gonalves et al. (2013)*).

From Table 9 we can see that our classifier has a much higher agreement with VADER than with spaCy (63% and 55% respectively). Pearson's rank correlation between our method and VADER came out to be 0.565, which indicates that they are sensibly correlated.

Table 7 Examples of sentiment classification results of VADER, spaCy, and PMI-based sentiment analysis for the manually annotated dataset containing 100 tweets.

Tweet	VADER Sent.	spaCy Sent.	PMI-based Sent.	PMI-based Score
Moral consistency requires veganism. Go vegan today!	Neutral	Neutral	Positive	0.164
I don't care what our desperate primate ancestors ate to *survive.* Real men do not murder weak, helpless, little creatures.	Positive	Negative	Neutral	0.009
Coke Bottle Redesign: Stackable, space friendly and 100% plant-based	Positive	Positive	Positive	1.096
A great review from MyBeautyBunny! Dr. Sharp's Cruelty-Free Dental Care Solution	Positive	Positive	Positive	1.462
Over 110M animals r killed in the food industry every single day. Can u really put off going vegan or talking 2 others about veganism?	Negative	Negative	Negative	−0.493
What's more like God? Wants you to enslave, breed and murder innocent beings? Or protect them from those acts?	Negative	Positive	Negative	−0.514

Table 8 Sentiment prediction accuracy for VADER, spaCy, PMI-based classifiers. We also provided accuracy for annotator A4, just for comparison.

	VADER Sent.	spaCy Sent.	PMI-based Sent.	A4 Sent.
Accuracy	0.69	0.67	0.77	0.87

Table 9 Percentage of agreement between our method, VADER, and spaCy.

	VADER Sent.	spaCy Sent.	PMI-based Sent.
VADER Sent.	–	69	63
spaCy Sent.	69	–	55
PMI-based Sent.	63	55	-

DISCUSSION

Let us consider how the current research results compare to other studies' results. Previously, there was similar research, but mostly on much broader topics, such as health, sports, and a healthy diet. We consider a narrower topic and study not only sentiments but also emotions for such a lengthy 12-year period. Such a study using Twitter data can confirm and supplement existing survey data on the subject and, in certain cases, even replace it. Our findings confirm previous studies that have reported a growth in the vegan trend (*Craig, 2009*; *López et al., 2019*; *Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*; *Cooper, 2018*). For example, according to the study by *Cooper (2018)*, climate catastrophe makes more people become vegan and reject animal speciesism.

Previous studies have shown that the number of people adopting a vegan or vegetarian diet is increasing (*López et al., 2019*). Another study revealed that Twitter users consider vegan, organic, and homemade foods the healthiest options (*Pila, Kvasnikov Stanislavsk & Kvasnika, 2021*). The other study by *Faunalytics (2021)* used Twitter to analyze how interest in the vegan diet has varied from 2019 to 2020. Our findings agree that vegan-related tweets

are more often used in a positive context than in a neutral or negative context. Next, the results of the study by *Lawo et al. (2020)* show that most people became vegan because they saw the reality of animal suffering on YouTube, Netflix documentary videos, and shows. Our findings confirm this, since we witnessed a clear spike in all emotions in the years of the documentaries release, especially '*Surprise*'.

According to research, a vegan diet is especially popular among teenagers and youth (*Paslakis et al., 2020*; *Craig, 2009*), so, probably, the drivers of the trend are primarily young people, who spend a lot of time on social networks (*Castillo de Mesa et al., 2020*).

We identified that the '*Fear*' emotions are also present. Moreover, they are growing, probably due to the risk of certain nutritional deficiencies that may be present if the vegan diet of an individual is not balanced (*Craig, 2009*) (not a whole food plant-based diet). Our findings support the earlier result obtained by *Twine (2017)*. They conducted a survey in the UK and identified that while transitioning to a vegan diet, people have fear because of a lack of nutritional knowledge. The same result was obtained by *Jennings et al. (2019)*. A vegan diet can indeed lead to deficiencies if it consists of French fries, bread, and vegan candy. These are vegan but unhealthy foods. A whole, plant-based diet rich in vegetables, fruits, whole grains, beans, legumes, seeds, and nuts needs to be promoted (*Willett et al., 2019*).

In addition, our results corroborate the findings of the study on the influence of sentiment on information diffusion (positivity bias) (*Ferrara & Yang, 2015*).

CONCLUSION

The vegan diet has been shown to provide a variety of health and environmental benefits in multiple studies. This study aimed to identify how the attitude toward a plant-based diet has been changing from 2010 to 2022 using Twitter social network data. We examined the prevailing sentiments and emotions embedded in tweets and found that veganism is currently frequently debated on social media compared to previous years. As the results suggest, there is a growing acceptance of veganism, the popularity of the vegan diet is expanding, and associated sentiments and emotions are becoming more positive. However, feelings of fear are also present.

Our findings place new information in the public domain, which has significant practical implications. Obtained results can be used to develop corresponding government action programs to promote healthy veganism and reduce the associated feeling of fear among the population. By gaining a deeper understanding of the public perception of veganism and using the substantial research unveiling the health benefits of the vegan diet (*Radnitz, Beezhold & DiMatteo, 2015*; *Willett et al., 2019*; *Norman & Klaus, 2019*; *Qian et al., 2019*; *Tran et al., 2020*; *Turner-McGrievy, Mandes & Crimarco, 2017*; *Park & Kim, 2022*), clinical practitioners and public health experts can develop efficient health programs and strategies and encourage more people to either follow a healthy vegan diet or reduce animal product consumption. In addition to health benefits, this can help to reduce the impact of climate change (*Dagevos & Voordouw, 2013*; *Willett et al., 2019*; *Scarborough et al., 2014*; *Masson-Delmotte et al., 2019*; *Cleveland & Gee, 2017*; *Chai et al., 2019*; *Eisen & Brown, 2022*). Even

Table 10 The main findings of the study.

Conclusion	Evidence and details
Sentiments and emotions in vegan-related tweets are becoming more positive.	See Fig. 15, Table 4, Figs. 16, 17
Positive sentiments in tweets are consistently dominating.	See Fig. 15, Table 4
Fear emotions were mostly medium and strong throughout the period.	See Fig. 16
There was a clear spike in all emotions in the years of the documentaries release, especially *Surprise*.	See Fig. 16
Sentiments started to grow during those years when the associated studies on the benefits of a plant-based diet appeared	See Fig. 15, Table 4
In 2022 the proportion of negative tweets relatively decreased compared to previous years. Possibly, due to the impact of COVID-19 and global warming, plant-based foods are becoming more common in post-pandemic diets.	See Fig. 15
Positive posts reach a wider audience (positive bias)	See Fig. 19
The PMI-based scoring method obtained higher accuracy than VADER and spaCY for vegan-specific tweets.	See Table 8

slight decreases in dairy and meat consumption by many people could make a significant contribution to climate change (*Scarborough et al., 2014*). Results also give us a sense of idea where this trend is heading because having a positive attitude is the first step toward moving on a plant-based diet. The main conclusions of our study are presented in Table 10.

The method we proposed was cross-validated on an annotated subsample of a dataset and achieved good accuracy (0.77). Our scoring method obtained much higher accuracy than VADER and spaCY for vegan-specific tweets. The main reason is that the dataset we used for training had narrow variability and context-specific vocabulary.

The study has certain limitations. Our analysis keywords are limited, but we plan to extend them in our future works. Next, the estimates of public perception of veganism were based on only one social network (Twitter); in future works, we plan to include other social media, namely Instagram and Facebook. We also plan to explore vegan tweets corresponding to the earlier period as well. Moreover, in this work, we focused on English tweets. However, the attitude toward a plant-based diet differs from country to country. Although English is popular in many countries, analyzing mainly English tweets may not accurately reflect the overall sentiment worldwide. For example, in Post-Soviet countries, acceptance and rise of veganism go much slower due to the absence of vegan communities and limited access to vegan products (*Aavik, 2019*). The popularity of avoiding or limiting meat products varies greatly by country, so it is important to consider different regions. Fellow researchers might now extend this study by exploring the public perception of veganism across different regions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Ministry of Education and Sciences of the Republic of Kazakhstan under the following grants: #AP09260670 ''Development of methods and algorithms for augmenting input data for modifying vector word embeddings'' and #AP14871214 ''Development of machine learning methods to increase the coherence of text in summaries produced by the Extractive Summarization Methods''. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Ministry of Education and Sciences of the Republic of Kazakhstan: #AP09260670, # AP14871214.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Elvina Shamoi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Akniyet Turdybay conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Pakizar Shamoi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Iskander Akhmetov performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, data scraping to prepare dataset, and approved the final draft.
- Assel Jaxylykova analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Alexandr Pak analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The annotated data is available in the Supplemental Files.

The data is available at GitHub: https://github.com/iskander-akhmetov/vegan_tweets_SA.

The data visualization is also available at GitHub: https://github.com/AlexPak/VeganTwitterEDA/blob/main/Visualization.ipynb.

The code is available at GitHub: https://github.com/elvinast/Sentiment-Analysis-Vegan-Tweets.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1149#supplemental-information.

REFERENCES

Aavik K. 2019. *The rise of veganism in post-socialist Europe: making sense of emergent vegan practices and identities in Estonia*. Reno: University of Nevada Press, 146–164 DOI 10.1093/isle/isaa007.

Adwan OY, Al-Tawil M, Huneiti AM, Shahin RA, Zayed AAA, Al-Dibsi RH. 2020. Twitter sentiment analysis approaches: a survey. *International Journal of Emerging Technologies in Learning* **15**:14467 DOI 10.3991/ijet.v15i15.14467.

Akhmetov I, Gelbukh A, Mussabayev R. 2022. Topic-aware sentiment analysis of news articles. *Computación y Sistemas* **26(1)**:4179 DOI 10.13053/cys-26-1-4179.

Akilandeswari J, Jothi G. 2018. Sentiment classification of tweets with non-language features. *Procedia Computer Science* **143**:426–433 8th international conference on advances in computing & communications (ICACC-2018) DOI 10.1016/j.procs.2018.10.414.

Alae-Carew C, Green R, Stewart C, Cook B, Dangour AD, Scheelbeek PF. 2022. The role of plant-based alternative foods in sustainable and healthy food systems: consumption trends in the UK. *Science of the Total Environment* **807**:151041 DOI 10.1016/j.scitotenv.2021.151041.

Alhgren M. 2022. Twitter statistics 2022: stats, user demography and facts. *Available at https://www.websiterating.com/research/twitter-statistics/*.

Aman F, Masood S. 2020. How nutrition can help to fight against COVID-19 pandemic: COVID-19 and diet. *Pakistan Journal of Medical Sciences* **36(COVID19-S4)**S121–S123 DOI 10.12669/pjms.36.COVID19-S4.2776.

Ashraf N, Khan L, Butt S, Chang H-T, Sidorov G, Gelbukh A. 2022. Multi-label emotion classification of Urdu tweets. *PeerJ Computer Science* **8**:e896 DOI 10.7717/peerj-cs.896.

Azizan A, Jamal NNSA, Abdullah MN, Mohamad M, Khairudin N. 2019. Lexicon-based sentiment analysis for movie review tweets. In: *2019 1st International conference on artificial intelligence and data sciences (AiDAS)*. 132–136 DOI 10.1109/AiDAS47888.2019.8970722.

Bindal N, Chatterjee N. 2016. A two-step method for sentiment analysis of tweets. In: *2016 International conference on information technology (ICIT)*. Piscataway: IEEE, 218–224 DOI 10.1109/ICIT.2016.052.

Bird SEL, Klein E. 2009. *Natural language processing with python*. Sebastopol: O'Reilly Media, Inc.

Bonzanini M. 2016. *Mastering social media mining with Python: acquire and analyze data from all corners of the social web with Python*. Birmingham: Packt Publishing.

Cambria E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* **31(2)**:102–107 DOI 10.1109/MIS.2016.31.

Castillo de Mesa J, Gmez-Jacinto L, Lpez Pelez A, Erro-Garcs A. 2020. Social networking sites and youth transition: the use of facebook and personal well-being of social work young graduates. *Frontiers in Psychology* **11**:230 DOI 10.3389/fpsyg.2020.00230.

Chai BC, van der Voort JR, Grofelnik K, Eliasdottir HG, Klss I, Perez-Cueto FJA. 2019. Which diet has the least environmental impact on our planet? A systematic review of vegan, vegetarian and omnivorous diets. *Sustainability* **11(15)**:4110 DOI 10.3390/su11154110.

Chinnalagu A, Durairaj AK. 2021. Context-based sentiment analysis on customer reviews using machine learning linear models. *PeerJ Computer Science* **7**:e813 DOI 10.7717/peerj-cs.813.

Church K, Hanks P. 2002. Word association norms, mutual information, and lexicography. *Computational Linguistics* **16(1)**:22–29 DOI 10.3115/981623.981633.

Cleveland DA, Gee Q. 2017. 9 - Plant-based diets for mitigating climate change. In: Mariotti F, ed. *Vegetarian and plant-based diets in health and disease prevention*. Cambridge: Academic Press, 135–156 DOI 10.1016/B978-0-12-803968-7.00009-5.

Cole M, Morgan K. 2011. Vegaphobia: derogatory Discourses of veganism and the reproduction of speciesism in UK national newspapers'. *British Journal of Sociology* **62(1)**:134–153 DOI 10.1111/j.1468-4446.2010.01348.x.

Cooper L. 2018. A new veganism: how climate change created more vegans. *Granite: Aberdeen University Postgraduate Interdisciplinary Journal* **2(1)**:16–24.

Craig WJ. 2009. Health effects of vegan diets. *The American Journal of Clinical Nutrition* **89(5)**:1627S–1633S DOI 10.3945/ajcn.2009.26736N.

Dagevos H, Voordouw J. 2013. Sustainability and meat consumption: is reduction realistic? *Sustainability: Science, Practice, and Policy* **9**:60–69 DOI 10.1080/15487733.2013.11908115.

Das S, Behera RK, kumar M, Rath SK. 2018. Real-time sentiment analysis of Twitter streaming data for stock prediction. *Procedia Computer Science* **132**:956–964 International Conference on Computational Intelligence and Data Science DOI 10.1016/j.procs.2018.05.111.

Di Renzo L, Gualtieri P, Pivari F, Soldati L, Attinà A, Cinelli G, Leggeri C, Caparello G, Barrea L, Scerbo F, Esposito E, De Lorenzo A. 2020. Eating habits and lifestyle changes during COVID-19 lockdown: an Italian survey. *Journal of Translational Medicine* **18(1)**:229 DOI 10.1186/s12967-020-02399-5.

Eisen MB, Brown PO. 2022. Rapid global phaseout of animal agriculture has the potential to stabilize greenhouse gas levels for 30 years and offset 68 percent of CO2 emissions. *PLOS Climate* **1(2)**:e0000010 DOI 10.1371/journal.pclm.0000010.

Fano R. 1961. *Transmission of information: a statistical theory of communication*. *MIT Press Classics*. Cambridge: MIT Press.

Faunalytics. 2021. Twitter Trends: #CageFree, #Vegan, #AnimalRights, and More! *PsyArXiv* DOI 10.31234/osf.io/gpcyd.

Feldman R. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* **56(4)**:82–89 DOI 10.1145/2436256.2436274.

Ferrara E, Yang Z. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* **1**:e26 DOI 10.7717/peerj-cs.26.

Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ, Singh GM, Gutierrez HR, Lu Y, Bahalim AN, Farzadfar F, Riley LM, Ezzati M. 2011. National, regional, and global trends in body-mass index since 1980: systematic

analysis of health examination surveys and epidemiological studies with 960 country-years and 9·1 million participants. *The Lancet* **377(9765)**:557–567 DOI 10.1016/S0140-6736(10)62037-5.

Ghosh A, Li G, Veale T, Rosso P, Shutova E, Reyes A, Barnden J. 2015. SemEval-2015 Task 11: sentiment analysis of figurative language in Twitter. In: *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015* DOI 10.18653/v1/S15-2080.

Gonalves P, Arajo M, Benevenuto F, Cha M. 2013. Comparing and combining sentiment analysis methods. In: *COSN 2013 - Proceedings of the 2013 conference on online social networks*. 27–38 DOI 10.1145/2512938.2512951.

Gore RJ, Diallo SPJ. 2015. You are what you tweet: connecting the geographic variation in americas obesity rate to Twitter content. *PLOS ONE* **10(9)**:e0133505 DOI 10.1371/journal.pone.0133505.

Hamdan H, Bellot P, Bechet F. 2015. Sentiment lexicon-based features for sentiment analysis in short text. *Research in Computing Science* **90**:217–226 DOI 10.13053/rcs-90-1-17.

Hemmatian F, Sohrabi MK. 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review* **52**:1495–1545 DOI 10.1007/S10462-017-9599-6/FIGURES/10.

Honnibal M, Montani I. 2017. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Available at https://spacy.io/*.

Hu M, Liu B. 2004. Mining and summarizing customer reviews. In: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04*. New York: Association for Computing Machinery, DOI 10.1145/1014052.1014073.

Hutto CJ, Gilbert E. 2014. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media. Available at https://ojs.aaai.org/index.php/ICWSM/article/view/14550*.

Islam T. 2019. Yoga-Veganism: correlation mining of twitter health data. In: *Conference: 8th KDD workshop on issues of sentiment discovery and opinion mining (WISDOM)* DOI 10.48550/arXiv.1906.07668.

James PT, Ali Z, Armitage AE, Bonell A, Cerami C, Drakesmith H, Jobe M, Jones KS, Liew Z, Moore SE, Morales-Berstein F, Nabwera HM, Nadjm B, Pasricha S-R, Scheelbeek P, Silver MJ, Teh MR, Prentice AM. 2021. The role of nutrition in COVID-19 susceptibility and severity of disease: a systematic review. *The Journal of Nutrition* **151(7)**:1854–1878 DOI 10.1093/jn/nxab059.

Jennings L, Danforth CM, Dodds PS, Pinel EC, Pope L. 2019. Exploring perceptions of veganism. ArXiv preprint. arXiv:1907.12567 DOI 10.48550/arXiv.1907.12567.

Jurafsky D, Martin J, Norvig P, Russell S. 2014. *Speech and language processing.* Boston, MA: Pearson Education.

JustAnotherArchivist. 2022. snscrape: a social networking service scraper in Python. Boston, MA: The GNU Press - Free Software Foundation office. *Available at https://github.com/JustAnotherArchivist/snscrape*.

Kanna PR, Pandiaraja P. 2019. An efficient sentiment analysis approach for product review using turney algorithm. *Procedia Computer Science* **165**:356–362 2nd International conference on recent trends in advanced computing ICRTAC -DISRUP - TIV INNOVATION, 2019 November 11-12, 2019 DOI 10.1016/j.procs.2020.01.038.

Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. 2018. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management* **38(1)**:1–6 DOI 10.1016/j.ijinfomgt.2017.08.002.

Karyotis C, Doctor F, Iqbal R, James A, Chang V. 2016. A fuzzy modelling approach of emotion for affective computing systems. In: *The first international conference on internet of things and big data, special session, recent advancement in internet of things, big data and security (RAIBS)* DOI 10.5220/0005945604530460.

Kelsingazin Y, Akhmetov I, Pak A. 2021. Sentiment analysis of kaspi product reviews. In: *2021 16th international conference on electronics computer and computation (ICECCO)*. 1–5 DOI 10.1109/ICECCO53203.2021.9663854.

Kim H, Rebholz CM, Hegde S, LaFiura C, Raghavan M, Lloyd JF, Cheng S, Seidelmann SB. 2021. Plant-based diets, pescatarian diets and COVID-19 severity: a population-based case—control study in six countries. *BMJ Nutrition, Prevention & Health* **4(1)**:257–266 DOI 10.1136/bmjnph-2021-000272.

Kirill Y, Mihail IG, Sanzhar M, Rustam M, Olga F, Ravil M. 2020. Propaganda identification using topic modelling. In: *Proceedings of the 9th international young scientist conference on computational science (YSC 2020), volume 178*. DOI 10.1016/j.procs.2020.11.022.

Kocaman V, Talby D. 2021. Spark NLP: natural language understanding at scale. *Software Impacts* **8**:100058 DOI 10.1016/j.simpa.2021.100058.

Lawo D, Esau M, Engelbutzeder P, Stevens G. 2020. Going Vegan: the Role(s) of ICT in vegan practice transformation. *Sustainability* **12(12)**:4997–5053 DOI 10.3390/su12125184.

Ligthart A, Catal C, Tekinerdogan B. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review* **54**:4997–5053 DOI 10.1007/s10462-021-09973-3.

López P, Cativo Calderon E, Atlas S, Rosendorff C. 2019. The effect of vegan diets on blood pressure in adults: a meta-analysis of randomized, controlled trials. *The American Journal of Medicine* **132(7)**:875–883.e7 DOI 10.1016/j.amjmed.2019.01.044.

MacInnis CC, Hodson G. 2017. It ain't easy eating greens: evidence of bias toward vegetarians and vegans from both source and target. *Group Processes & Intergroup Relations* **20(6)**:721–744 DOI 10.1177/1368430215618253.

Masson-Delmotte V, Zhai P, Prtner HO, Roberts D, Skea J, Shukla P, Pirani A, Moufouma-Okia W, Pan C, Pidcock R, Connors S, Matthews JBR, Chen Y, Zhou X, Gomis MI, Lonnoy E, Maycock T, Tignor M, Waterfield T. 2019. UN Intergovernmental Panel on Climate Change (IPCC), 2018: global warming of 1.5°C.

Available at https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_Low_Res.pdf .

Matiko JW, Beeby SP, Tudor J. 2014. Fuzzy logic based emotion classification. In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Piscataway: IEEE, 4389–4393 DOI 10.1109/ICASSP.2014.6854431.

Merino J, Joshi AD, Nguyen LH, Leeming ER, Mazidi M, Drew DA, Gibson R, Graham MS, Lo C-H, Capdevila J, Murray B, Hu C, Selvachandran S, Hammers A, Bhupathiraju SN, Sharma SV, Sudre C, Astley CM, Chavarro JE, Kwon S, Ma W, Menni C, Willett WC, Ourselin S, Steves CJ, Wolf J, Franks PW, Spector TD, Berry S, Chan AT. 2021. Diet quality and risk and severity of COVID-19: a prospective cohort study. *Gut* **70(11)**:2096–2104 DOI 10.1136/gutjnl-2021-325353.

Mukhamediev RI, Yakunin K, Mussabayev R., Buldybayev T, Kuchin Y, Murzakhmetov S, Yelis M. 2020. Classification of negative information on socially significant topics in mass media. *Symmetry* **12(12)**:1945 DOI 10.3390/sym12121945.

Neethu MS, Rajasree R. 2013. Sentiment analysis in Twitter using machine learning techniques. In: *2013 Fourth international conference on computing, communications and networking technologies (ICCCNT)*. 1–5 DOI 10.1109/ICCCNT.2013.6726818.

Norman K, Klaus S. 2019. Veganism, aging and longevity: new insight into old concepts. *Current Opinion in Clinical Nutrition and Metabolic Care* **23**:145–150 DOI 10.1097/MCO.0000000000000625.

Omarkhan M, Kissymova G, Akhmetov I. 2021. Handling data imbalance using CNN and LSTM in financial news sentiment analysis. In: *2021 16th international conference on electronics computer and computation (ICECCO)*. 1–8 DOI 10.1109/ICECCO53203.2021.9663802.

Ostic D, Qalati SA, Barbosa B, Shah SMM, Galvan Vela E, Herzallah AM, Liu F. 2021. Effects of social media use on psychological well-being: a mediated model. *Frontiers in Psychology* **12**:678766 DOI 10.3389/fpsyg.2021.678766.

Park E, Kim S-B. 2022. Veganism during the COVID-19 pandemic: vegans' and nonvegans' perspectives. *Appetite* **175**:106082 DOI 10.1016/j.appet.2022.106082.

Paslakis G, Richardson C, Nöhre M, Brähler E, Holzapfel C, Hilbert A, de Zwaan M. 2020. Prevalence and psychopathology of vegetarians and vegans—results from a representative survey in Germany. *Scientific Reports* **10(1)**:6840 DOI 10.1038/s41598-020-63910-y.

Pila L, Kvasnikov Stanislavsk L, Kvasnika R. 2021. Healthy food on the Twitter social network: vegan, homemade, and organic food. *International Journal of Environmental Research and Public Health* **18(7)** DOI 10.3390/ijerph18073815.

Python Software Foundation. 2020. text2emotion 0.0.5 package. *Available at https://pypi.org/project/text2emotion/*.

Qamar S, Ahmad P. 2015. Emotion detection from text using fuzzy logic. *International Journal of Computer Applications* **121**:29–32 DOI 10.5120/21522-4501.

Qian F, Liu G, Hu F, Bhupathiraju S, Sun Q. 2019. Association between plant-based dietary patterns and risk of type 2 diabetes: a systematic review and meta-analysis. *JAMA Internal Medicine* **179(10)**:1335–1344 DOI 10.1001/jamainternmed.2019.2195.

Radnitz C, Beezhold B, DiMatteo J. 2015. Investigation of lifestyle choices of individuals following a vegan diet for health and ethical reasons. *Appetite* **90**:31–36 DOI 10.1016/j.appet.2015.02.026.

Revels S, Kumar SA, Ben-Assuli O. 2017. Predicting obesity rate and obesity-related healthcare costs using data analytics. *Health Policy and Technology* **6(2)**:198–207 DOI 10.1016/j.hlpt.2017.02.002.

Reyes-Menendez A, Saura JR, Alvarez-Alonso C. 2018. Understanding worldenvironmentday user opinions in Twitter: a topic-based sentiment analysis approach. *International Journal of Environmental Research and Public Health* **15(11)**:2537 DOI 10.3390/ijerph15112537.

Reyes-Menendez A, Saura JR, Filipe F. 2020. Marketing challenges in the MeToo era: gaining business insights using an exploratory sentiment analysis. *Heliyon* **6(3)**:e03626 DOI 10.1016/j.heliyon.2020.e03626.

Rzymski P, Kulus M, Jankowski M, Dompe C, Bryl R, Petitte JN, Kempisty B, Mozdziak P. 2021. COVID-19 Pandemic is a call to search for alternative protein sources as food and feed: a review of possibilities. *Nutrients* **13(1)**:150 DOI 10.3390/nu13010150.

Sailunaz K, Alhajj R. 2019. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science* **36**:101003 DOI 10.1016/j.jocs.2019.05.009.

Saura JR, Palos-Sanchez P, Rios Martin MA. 2018. Attitudes expressed in online comments about environmental factors in the tourism sector: an exploratory study. *International Journal of Environmental Research and Public Health* **15(3)**:553 DOI 10.3390/ijerph15030553.

Saura JR, Reyes-Menendez A, Palos-Sanchez P. 2019. Are black friday deals worth it? Mining Twitter users sentiment and behavior response. *Journal of Open Innovation: Technology, Market, and Complexity* **5(3)**:58 DOI 10.3390/joitmc5030058.

Scarborough P, Appleby P, Mizdrak A, Briggs A, Travis R, Bradbury K, Key T. 2014. Dietary greenhouse gas emissions of meat-eaters, fish-eaters, vegetarians and vegans in the UK. *Climatic Change* **125**:179–192 DOI 10.1007/s10584-014-1169-1.

Serrano L, Ariza-Montes A, Nader M, Sianes A, Law R. 2021. Exploring preferences and sustainable attitudes of Airbnb green users in the review comments and ratings: a text mining approach. *Journal of Sustainable Tourism* **29(7)**:1134–1152 DOI 10.1080/09669582.2020.1838529.

Shamoi P, Inoue A, Kawanaka H. 2016. FHSI: toward more human-consistent color representation. *Journal of Advanced Computational Intelligence and Intelligent Informatics* **20(3)**:393–401.

Shamoi P, Inoue A. 2012. Computing with words for direct marketing support system. In: *Midwest artificial intelligence and cognitive science conference*.

Shamoi P, Inoue A, Kawanaka H. 2014. Fuzzy color space for apparel coordination. *Open Journal of Information Systems (OJIS)* 20–28 *Available at http://nbn-resolving.de/urn:nbn:de:101:1-201705194710*.

Shamoi P, Inoue A, Kawanaka H. 2020. Modeling aesthetic preferences: color coordination and fuzzy sets. *Fuzzy Sets and Systems* **395**:217–234 DOI 10.1016/j.fss.2019.02.014.

Sims S. 2015. Sentiment analysis 101 - KDnuggets. *Available at https://www.kdnuggets.com/2015/12/sentiment-analysis-101.html#:~:text=Sentiment%20analysis%20is%20a%20term,writer%20toward%20a%20specific%20topic*.

Singh B, Jain S, Rastogi A. 2021. Effects of nationwide COVID-19 lockdown on lifestyle and diet: an Indian survey. *Journal of Family Medicine and Primary Care* **10(3)**1246–1250 DOI 10.4103/jfmpc.jfmpc_2046_20.

Tran E, Dale H, Jensen C, Lied G. 2020. Effects of plant-based diets on weight status: a systematic review. *Diabetes, Metabolic Syndrome and Obesity* **13**:3433–3448 DOI 10.2147/DMSO.S272802.

Turner-McGrievy G, Mandes T, Crimarco A. 2017. A plant-based diet for overweight and obesity prevention and treatment. *Journal of Geriatric Cardiology* **14**:369–374 DOI 10.11909/j.issn.1671-5411.2017.05.002.

Turney P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. DOI 10.3115/1073083.1073153.

Turney P, Littman M. 2003a. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems* **21**:315 DOI 10.1145/944012.944013.

Turney P, Littman M. 2003b. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *ArXiv* DOI 10.48550/arXiv.cs/0212012.

Twine R. 2017. A practice theory framework for understanding vegan transition. *Animal Studies Journal* **6(2)**:192–224.

Utama H. 2019. Sentiment analysis in airline tweets using mutual information for feature selection. In: *2019 4th international conference on information technology, information systems and electrical engineering (ICITISEE)*. 295–300 DOI 10.1109/ICITISEE48480.2019.9003903.

Vo A-D, Ock C-Y. 2012. Sentiment classification: a combination of PMI, sentiWordNet and fuzzy function. In: *International conference on computational collective intelligence*. 373–382 DOI 10.1007/978-3-642-34707-8_38.

WHO WHO. 2021. Plant-based diets and their impact on health, sustainability and the environment: a review of the evidence: WHO European Office for the Prevention and Control of Noncommunicable Diseases. Technical documents. World Health Organization. Regional Office for Europe 11 p..

Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, Garnett T, Tilman D, DeClerck F, Wood A, Jonell M, Clark M, Gordon LJ, Fanzo J, Hawkes C, Zurayk R, Rivera JA, De Vries W, Majele Sibanda L, Afshin A, Chaudhary A, Herrero M, Agustina R, Branca F, Lartey A, Fan S, Crona B, Fox E, Bignet V, Troell M, Lindahl T, Singh S, Cornell SE, Srinath Reddy K, Narain S, Nishtar S, Murray CJL. 2019. Food in the anthropocene: the EAT lancet commission on healthy diets from sustainable food systems. *The Lancet* **393(10170)**:447–492 DOI 10.1016/S0140-6736(18)31788-4.

Williams S, Terras M, Warwick C. 2013. What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation* **69**:384–410 DOI 10.1108/JD-03-2012-0027.

Zabetakis I, Matthys C, Tsoupras A. 2021. Editorial: Coronavirus Disease (COVID-19): diet, inflammation and nutritional status. *Frontiers in Nutrition* **8**:760720 DOI 10.3389/fnut.2021.760720.

Zadeh L. 1975. The concept of a linguistic variable and its application to approximate reasoning-III. *Information Sciences* **9(1)**:43–80 DOI 10.1016/0020-0255(75)90017-1.

Zadeh LA. 1965. Fuzzy sets. *Information and Control* **8**:338–353 DOI 10.1016/S0019-9958(65)90241-X.

Zubiaga A, Spina D, Fernndez V, Martnez-Unanue R. 2014. Real-time classification of Twitter trends. *Journal of the American Society for Information Science and Technology* **66**:462–473 DOI 10.1002/asi.23186.

PeerJ Computer Science

Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations

Rasha Obeidat, Maram Gharaibeh, Malak Abdullah and Yara Alharahsheh

Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

ABSTRACT

Since the inception of the current COVID-19 pandemic, related misleading information has spread at a remarkable rate on social media, leading to serious implications for individuals and societies. Although COVID-19 looks to be ending for most places after the sharp shock of Omicron, severe new variants can emerge and cause new waves, especially if the variants can evade the insufficient immunity provided by prior infection and incomplete vaccination. Fighting the fake news that promotes vaccine hesitancy, for instance, is crucial for the success of the global vaccination programs and thus achieving herd immunity. To combat the proliferation of COVID-19-related misinformation, considerable research efforts have been and are still being dedicated to building and sharing COVID-19 misinformation detection datasets and models for Arabic and other languages. However, most of these datasets provide binary (true/false) misinformation classifications. Besides, the few studies that support multi-class misinformation classification deal with a small set of misinformation classes or mix them with situational information classes. False news stories about COVID-19 are not equal; some tend to have more sinister effects than others (*e.g.*, fake cures and false vaccine info). This suggests that identifying the sub-type of misinformation is critical for choosing the suitable action based on their level of seriousness, ranging from assigning warning labels to the susceptible post to removing the misleading post instantly. We develop comprehensive annotation guidelines in this work that define 19 fine-grained misinformation classes. Then, we release the first Arabic COVID-19-related misinformation dataset comprising about 6.7K tweets with multi-class and multi-label misinformation annotations. In addition, we release a version of the dataset to be the first Twitter Arabic dataset annotated exclusively with six different situational information classes. Identifying situational information (*e.g.*, caution, help-seeking) helps authorities or individuals understand the situation during emergencies. To confirm the validity of the collected data, we define three classification tasks and experiment with various machine learning and transformer-based classifiers to offer baseline results for future research. The experimental results indicate the quality and validity of the data and its suitability for constructing misinformation and situational information classification models. The results also demonstrate the superiority of AraBERT-COV19, a transformer-based model pretrained on COVID-19-related tweets, with micro-averaged F-scores of 81.6% and 78.8% for the multi-class misinformation and situational information classification

Submitted 30 May 2022
Accepted 17 October 2022
Published 5 December 2022

Corresponding authors
Rasha Obeidat,
rmobeidat@just.edu.jo
Malak Abdullah,
mabdullah@just.edu.jo

Academic editor
Ana Maguitman

Additional Information and Declarations can be found on page 29

DOI **10.7717/peerj-cs.1151**

Copyright
2022 Obeidat et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

tasks, respectively. Label Powerset with linear SVC achieved the best performance among the presented methods for multi-label misinformation classification with micro-averaged F-scores of 76.69%.

Subjects Data Mining and Machine Learning, Natural Language and Speech, Network Science and Online Social Networks, Social Computing, Neural Networks
Keywords COVID-19, Misinformation detection, Situational information, Deep learning, Transformers, BERT, Machine learning, Fake news, Data annotation, Data collection

INTRODUCTION

The emergence of the Coronavirus pandemic 2019 (COVID-19) has not only posed major challenges to health systems all over the globe but also has flooded social media platforms with excessive amounts of information about the disease mixed with misleading information, which is being referred to as the global infodemic. In the Arab region, Twitter has become one of the main communication channels for people to share news and opinions about COVID-19 (*Essam & Abdo, 2021*). This has been accompanied by widespread conspiracy theories, fake cures, and anti-vaccine propaganda, making it very challenging to verify the shared information's truth. A study by *Khalifa et al. (2020)*, involving participants from nine Arab countries, shows that 33.52% of respondents believe that Twitter facilitates the spreading of coronavirus-related rumors. Another study that analyzes the magnitude of misinformation being spread on Twitter on COVID-19 reveals that 42.2% of the investigated tweets included misleading or unverifiable information (*Kouzy et al., 2020*). Further exacerbating the problem, fake news propagates on social media more rapidly than reliable news (*Dizikes, 2018*).

With the emergence of novel diseases that are not fully understood yet, health misinformation on social media can promote risky practices, posing serious threats to public health. Myriad false claims are rampant online about how to cure coronavirus and prevent its transmission without medical evidence. Whereas some claims, such as "eating garlic and drinking hot water can eliminate the virus," are harmless, some remedies can be life-threatening. Several fatalities and hospitalizations have been reported globally due to promoting the consumption of toxic substances to cure the virus (*Barua et al., 2020*; *Islam et al., 2020*). Besides, several conspiracy theories have gone viral on social media, such as the idea that 5G mobile networks are linked with the virus (*Alqurashi et al., 2021*) and that the virus is a biological weapon (*Matt Burgess, 2021*). Exposure to COVID-19 conspiracies reduces the effectiveness of official attempts to control the crisis and causes mental health issues that range from anxiety and depression to experiencing suicidal thoughts (*Su et al., 2021*).

At this stage of the pandemic, the harmful implications of COVID-19 misinformation are manifested in vaccine hesitancy. It is estimated that COVID-19 vaccines must be accepted by at least 55% of the population to obtain herd immunity (*Loomba et al., 2021*). However, achieving high public uptake of vaccines is threatened by conspiracy theories and mistrust (*Simione et al., 2021*). A relevant study (*AlTarrah et al., 2021*) that assessed

the public perceptions of the vaccine in Jordan and other Arab countries found a strong link between vaccine hesitancy and conspiratorial beliefs. For instance, 27.7% and 23.4% of respondents think that COVID-19 vaccines are intended to inject microchips into recipients and hurt fertility, respectively. Subsequently, fighting COVID-19 misinformation has become more important than ever to achieve a widespread acceptance of the vaccines for successful control of the pandemic.

Public health officials such as The World Health Organization (WHO) teamed up with leading social media companies, and fact-checking Organizations have made considerable efforts to mitigate the adverse consequences of misinformation and disseminate timely and evidence-based information to the public. However, manual fact-checking is time-consuming and may need domain expertise (*Greene & Murphy, 2021*). Thus, there is an urgent need to develop automated models that efficiently detect misinformation about COVID-19 on social media. Since the quality of misinformation detection models depends heavily on the availability of labeled data, considerable efforts have been made to create annotated datasets for COVID-19 misinformation detection from Arabic content (especially Twitter). Most of these works have focused exclusively on assigning binary labels ("true" and "false") to news stories (*Haouari et al., 2020b*; *Alqurashi et al., 2021*; *Mahlous & Al-Laith, 2021*). However, misleading stories about COVID-19 are not equal; they have varying levels of potential impacts on individuals and society (*Kabha et al., 2020*). Some tend to have more baleful than others (*e.g.*, fake cures). This suggests that identifying the sub-type of misinformation is critical for choosing the suitable action to deal with these types based on their level of seriousness.

To this end, we present and publicly share a misinformation dataset of 6.7K Arabic tweets spreading COVID-19-related claims. Our dataset is manually labeled for 19 misinformation classes representing the sub-types of the veracity (*i.e.*, falsity/truthfulness) of the information on Twitter. "Fake cure", "true cure," "fake vaccine info", and "true vaccine info" are examples of the defined misinformation classes. For instance, the tweet "تحذير! أكثر من 31 طبيبا وعالما يقولون ان لقاح كوفيد 19 ليس آمنًا وغير فعال!" (Translation: Warning! more than 31 doctors and scientists say that the COVID-19 vaccine is unsafe and ineffective!) is classified as "fake vaccine info," whereas we assigned the class "true vaccine info" to the tweet "الام ار ان اي لا يدخل النواة أصلا لكي يتلاعب في الجينات" (Translation: The mRNA does not enter the nucleus at all to manipulate the genes). In addition, we noticed that 8.29% of the tweets belong to more than one sub-type, wherefore we present a multi-label version of our dataset in which each tweet receives all of the relevant misinformation classes besides the assigned primary class. We also inspected different situational information types in Arabic tweets (*e.g.*, "caution and advice"). Since understanding situational information helps authorities or individuals understand the situation during emergencies (*Li et al., 2020a*), we present another version of our dataset annotated with situational information classes similar to the classes defined by *Li et al. (2020a)*. Besides, we design several learning tasks and experiment with various classifiers to offer baseline results for future research. Finally, we present a discussion of the potential applications of the proposed work. We believe that the misinformation detection systems that support fine-grained misinformation and situational classes are more capable of serving the specific

goals of the potentially interested parties, such as social media platforms and health professionals, compared binary classification. The contributions of this work are summarized as follows:

- We developed comprehensive misinformation annotation guidelines and released the first COVID-19-related Arabic dataset comprising about 6.7K tweets annotated with fine-grained multi-class and multi-label misinformation classes, including tweets related to COVID-19 vaccination.
- We develop comprehensive situational information annotation guidelines, and we release a version of our dataset to be the first tweets Arabic dataset annotated exclusively with situational information classes.
- We define several learning tasks over annotated data and present the experimental results of various transformers-based modes and ML classifiers using multiple feature extraction techniques.

PRELIMINARIES

In recent years, there has been a growing interest in detecting "fake news," "misinformation," and "disinformation" in social media. However, researchers have no agreement on the definition of these overlapping terms (*Zhou & Zafarani, 2020*; *Patwa et al., 2021*). One definition of fake news adopted in various recent research works is "a claim or piece of information that is verified to be false or misleading" (*Alam et al., 2020*; *Wani et al., 2021*). Several studies further divide fake news into "misinformation" and "disinformation" based on the author's intention (*Ciampaglia et al., 2018*; *Raza & Ding, 2022*; *Kolluri & Murthy, 2021*). Whereas "misinformation" is the unintentional sharing of false information, "disinformation" is the deliberate creation and sharing of false information with malicious intent to deceive people (*Ciampaglia et al., 2018*; *Kim et al., 2020*).

Fine-grained definitions of related information disorders are also used in the literature. Examples include mal-information, news fabrications, rumors, hoaxes, satire and propaganda. Distinguishing these types of information disorders is performed based on three dimensions: the factuality of information, the intent to harm and whether the reader is implicitly aware that the information could be false or not. Mal-information is the sharing of genuine information exaggerated with the intent to inflict harm (*Alonso et al., 2021*). News fabrications are stories about non-existing events such as celebrity gossip (*Pérez-Rosas et al., 2017*). Hoaxes are intentionally crafted fake stories that aim to convince people of their veracity (*Tacchini et al., 2017*). *Satire* is a humorous news item with genuine content intending to dispense irony (*De Sarkar, Yang & Mukherjee, 2018*). Rumor is pieces of information that are unverified at the time of posting and can be verified later; in other words, they can't be opinions or feelings (*Oshikawa, Qian & Wang, 2020*). Propaganda is a biased or exaggerated story that aims to manipulate readers to advance a specific agenda (*Da San Martino et al., 2019*; *Abdullah, Altiti & Obiedat, 2022*).

Since it is extremely hard to know the intention of publishing false information about COVID-19 in this work, we follow several previous studies (*Alqurashi et al., 2021*;

Pérez-Rosas et al., 2017; *De Sarkar, Yang & Mukherjee, 2018*; *Hossain, 2021*) and use the term "misinformation" as an umbrella term for all false and inaccurate information. We also use the term "fake news" as an interchangeable term, especially in the related work section. For a detailed overview of the fundamental theories related to different types of information disorders, the reader is highly encouraged to refer to the survey by *Zhou & Zafarani (2020)*.

RELATED WORK

Since the start of the COVID-19 outbreak, the study of COVID-19 misinformation has become a popular area of research. This section highlights some datasets that have been released to analyze social media data related to the COVID-19 pandemic in general. After that, we focus on the most relevant studies on misinformation detection. Then, we overview the main models and approaches proposed for COVID-19 misinformation detection from social media for the Arabic language.

COVID-19 datasets

A considerable number of COVID-19 datasets have emerged in a short span of time after the pandemic started with multiple different goals. *Mega-COV* (*Abdul-Mageed et al., 2020*) is a billion-scale multilingual COVID dataset collected from Twitter to study a wide host of phenomena related to the pandemic. It covers 65 languages and 268 countries and has more than 169M location-tagged tweets. *Geocov19* (*Qazi, Imran & Ofli, 2020*) is a multilingual COVID-19 dataset of hundreds of millions of tweets with location information. *CORD-19* (*Wang et al., 2020*) is a growing collection of full-text scientific papers regarding COVID-19, along with their associated metadata. The Allen Institute has designed the dataset to support literature-based discoveries. *Chen, Lerman & Ferrara (2020)* collected a multilingual COVID-19 Twitter data set to understand dynamics observable in social networks during the pandemic and track the spreading of COVID-19 information on Twitter.

Several datasets for analyzing the public responses to the pandemic have been made available. *Li et al. (2020a)* studied different types of situational information about COVID-19 (*e.g.*, "caution and advice") and its propagation on Weibo. *Banda et al. (2021)* collected a large-scale English dataset of over 1.12 billion tweets related to COVID-19 and made it freely available to support a wide range of research, such as epidemiological analyses and emotional and mental responses to official measures. *Kleinberg, van der Vegt & Mozes (2020)* released a dataset of 5,000 texts regarding COVID-19 with ground truth emotional responses. *Medford et al. (2020)* built a dataset to identify the sentiment polarity and predominant emotions that appeared on Twitter during the early stages of the pandemic. *Gupta, Vishwanath & Yang (2020)* created a COVID-19 dataset comprising over 198M tweets, labeled automatically using ML and NLP techniques with latent topics, sentiment, and emotions.

Similar datasets have been released for Arabic as well. *ArCOV-19* (*Haouari et al., 2020a*) is a geo-tagged COVID-19 Twitter dataset that comprises about 2.7M tweets collected *via* Twitter search API and covers a whole year of the pandemic. *Yang et al. (2020)* created an

English and Arabic Twitter dataset of 10K tweets annotated for the task of fine-grained sentiment analysis. *Alsudias & Rayson (2020)* collected 1M Arabic tweets related to COVID-19 and clustered them using the K-means algorithm into five clusters: statistics, prayers, disease locations, advising, and advertising. *Alomari et al. (2021)* released a dataset comprising 14 million tweets from the Kingdom of Saudi Arabia (KSA) and used unsupervised Latent Dirichlet Allocation (LDA) to detect 15 government pandemic measures and public concerns.

Although these datasets have been examined and contributed to understanding social media dynamics during the pandemic, they do not provide ground truth annotation of the truthfulness of coronavirus-related information to help fight the accompanying infodemic.

COVID-19 misinformation dataset

Many researchers have created and released COVID-19 misinformation detection datasets; the majority are with binary annotations. *Patwa et al. (2021)* manually annotated a COVID-19 fake news detection dataset containing 10,700 English social media posts and articles with binary labels ("real" and "fake"). *COVIDLIES* (*Hossain et al., 2020*) is a binary misinformation dataset containing 6,761 annotated English tweets. *Helmstetter & Paulheim (2021)* created a dataset of COVID-19 tweets for English classified into fake and non-fake tweets based on the trustworthiness of their source (*i.e.*, weak supervision). Likewise, *Zhou et al. (2020)* designed *ReCOVery*, a multimodal dataset of COVID-19 news articles and related tweets annotated weakly for veracity. Instead of fact-checking the news pieces separately, the veracity is decided collectively based on the credibility of the new sites. *CoAID* (*Cui & Lee, 2020*) is a diverse COVID-19 healthcare misinformation dataset, including fake claims crawled from reliable media outlets and automatically extracted tweets and replies related to these claims. *MM-COVID* (*Li et al., 2020b*) is a multilingual and multimodal dataset containing binary classified news content with the associated social engagements and spatial-temporal information for six different languages.

For the Arabic language, *Haouari et al. (2020b)* created *ArCOV19-Rumors*, an Arabic COVID-19 misinformation Twitter dataset containing 9.4K tweets, manually annotated into "false," "true," or "other" depending on the popular fact-checking websites. *Alqurashi et al. (2021)* created a large Arabic Twitter dataset related to the COVID-19 pandemic. The tweets are manually classified as tweets with misinformation ("false") and tweets with not-misleading information ("true"). *Elhadad, Li & Gebali (2020)* published *COVID-19-FAKES*, an English and Arabic COVID-19 tweets dataset, annotated automatically by using a community of 13 ML algorithms. *Mahlous & Al-Laith (2021)* collected more than seven million coronavirus-related Arabic tweets using trending hashtags and relied on France-Press Agency and the Saudi Anti-Rumors Authority. They manually fact-checked a sample of 2,500 tweets and annotated them into "fake" or "genuine" classes. CLEF2020 and CLEF2021 CheckThat! Lab shared tasks (*Nakov et al., 2021*; *Barrón-Cedeño et al., 2020*) released data covering several sub-tasks related to the factuality of Twitter posts on various topics, including COVID-19. For instance, one task is predicting whether a tweet is worth manual fact-checking or not (*i.e.*, the tweet is assigned to one of the classes "worth fact-checking" or "doesn't worth fact-checking"). Another task is predicting the veracity of a

news article and its topics in several languages, including Arabic (*Shahi, Struß & Mandl, 2021*).

Class diversity

In terms of class diversity, few works provide multi-class or multi-label classification. For example, *FakeCovid* (*Shahi & Nandini, 2020*) is a multilingual cross-domain dataset of news articles for COVID-19 that are manually fact-checked from 92 different fact-checking websites and labels with 11 categories including "true," "mostly true," "partially true," "false" *etc*. *CMU-MisCOV19* (*Memon & Carley, 2020*) is an English multi-class misinformation tweets dataset containing fact-checked claims manually annotated with 17 fine-grained misinformation labels. To our knowledge, *Ameur & Aliane (2021)*, *Mubarak & Hassan (2020)* and *Alam et al. (2020)* are the only studies that released Arabic misinformation datasets with multi-class or multi-label annotations. These studies are the closest to our work; however, none of them offers multi-label multi-class fine-grained misinformation classification.

Specifically, *Mubarak & Hassan (2020)* manually annotated a dataset of Arabic tweets collected for COVID-19, labeled for 13 classes, including one misinformation-related class (*i.e.*, "rumor"), COVID-related topics (*e.g.*, "cure," "virus info," "governmental measures,") and situational information classes (*e.g.*, "advice," "support," *etc.*). However, this annotation scheme suffers from two problems. First, it combines rumor-related tweet identification and situational information classification in one scheme, although they serve different purposes. Second, the scheme covers important topics of the pandemic that interest people but doesn't evaluate its veracity. For example, recognizing whether this cure is fake or real is more important than only identifying that a tweet promotes a cure. Third, it doesn't support multi-label classification; thus, If there is a rumor about some governmental measure, it is unclear whether it should be labeled as a "rumor" or a "governmental measure." This work solved these issues by providing multiple versions of our dataset. The first two versions are annotated with 19 fine-grained multi-label and multi-class misinformation annotations (*e.g.*, "true treatment or cure," "fake cure," "true gov measure," *etc.*), and a third version is annotated solely with six situational information classes.

Ameur & Aliane (2021) created *ARACOVID19-MFH*, a dataset that combines ten independent sub-tasks (each with predefined mutually exclusive labels) in one task. For example, The subtask "factual" specify the factuality of the tweet (*i.e.*, whether the tweet describes "real" or "fake" news), the subtask "contains hate" indicates whether the label contains hate speech or not, another subtask specifies the "dialect" of the tweet, and so on. Likewise, *Alam et al. (2020)* designed dataset for a multi-label Twitter dataset for COVID-19 misinformation analysis that covers the English and Arabic languages. The annotation is performed by answering seven independent questions with predefined answers (labels) about the various aspects of the tweets, such as the worthiness of fact-checking, the need for authorities' attention, *etc*. While the suggested annotation schemes assess different independent aspects of the data, they do not present a typical multi-label setting in which tweets receive one or more mutually non-exclusive misinformation

sub-type. By analyzing COVID-19-related Twitter data, We found that the same tweet could be assigned one or more veracity sub-type based on the contents. For example, one tweet could contain "fake cure" and "fake symptoms" simultaneously. To this end, we released a version of our dataset in which tweets receive all the applicable misinformation classes. To the best of our knowledge, our dataset is the first Arabic COVID-19 dataset that supports multi-label multi-class misinformation classification. Table 1 summarizes the existing COVID-19 datasets collected for misinformation detection or situational information classification including our dataset versions, including the size, the language, the sources of samples, the annotation method, and the class granularity.

Existing models and approaches

There has been substantial work in leveraging machine learning (ML) and deep learning (DL) to detect and classify COVID-19 misinformation. *Alsudias & Rayson (2020)* applied several ML algorithms using a sample of 2,000 tweets labeled manually for misinformation. logistic regression (LR), support vector classification (SVC), and naive Bayes (NB) classifiers are trained with two sets of features (TF-IDF and, FastText, and word2vec embeddings) to identify the rumor-related tweets with 84% accuracy achieved by LR. *Haouari et al. (2020b)* experimented with variant DL models to perform claim-level and tweet-level veracity verification using the ArCOV19-Rumors dataset they collected. The considered models are bidirectional graph convolutional network (Bi-GCN), PPC-RNN+CNN, and MARBERT and AraBERT transformer models. PPC-RNN+CNN *Liu & Wu (2018)* is a time series model that combines convolutional neural networks (CNN) and recurrent neural networks (RNN) with the propagation path classification (PPC). PPC is an algorithm that exploits user profiles to verify a tweet's genuineness. The results show that MARBERT achieved the best performance with a macro-F1 score of 74%. *Mubarak & Hassan (2020)* tested the performance of AraBERT transformer models and Support Vector Machine (SVM) classifier with several features, including word n-grams, character n-grams and Mazajak Embeddings (*Farha & Magdy, 2019*). The models are trained and tested using the multi-class *ArCorona* dataset they built. The experiments prove the superiority of AraBERT over SVM models with a macro-averaged F-score of 60.5%.

Ameur & Aliane (2021) evaluate the performance of five transformer models they provide as baselines for the ten independent tasks in the *ARACOVID19-MFH* dataset, including the binary misinformation detection task. The tested models are: AraBERT, mBERT, DistilBERT, AraBERT-COV19, and mBERT-COV19. The experimental results show that the last two transformer models, which have been fine-tuned using COVID-19 data, achieve the best misinformation detection performance with weighted F-scores of 95.78% and 94.91%, respectively. Likewise, *Alam et al. (2020)* demonstrate the transformer models' efficacy in performing binary and multi-class misinformation detection and fact-checking worthiness tasks. *Alqurashi et al. (2021)* applied eight different deep neural networks and ML models with multiple features, including FASTTEXT and word2ve word embeddings and word frequency, to detect COVID-19 misinformation in Arabic Twitter data. Experiments show that Extreme Gradient Boosting (XGBoost) is the best among the

Table 1 A summary of the existing COVID-19 datasets collected for misinformation detection or situational information classification. The summary includes the size, the language, the sources of samples, the annotation method, and the class granularity. All listed datasets are designed for misinformation detection except the ones released for situational info classification.

Dataset	Size	Main input	Language	Classification task	Annotation method
Patwa et al. (2021)	10.7K	Twitter, Facebook Fact-checking websites	English	Binary (fake, real)	Manually + fact-checked claims
COVIDLies (*Hossain et al., 2020*)	6.7K	Twitter	English	Binary (fake, real)	Manually
Helmstetter & Paulheim (2021)	400K	Twitter	English	Binary (fake, non-fake)	Weak supervision
ReCOVery (*Zhou et al., 2020*)	146K	News articles Twitter	English	Binary (fake, real)	Distance Supervision
CoAID (*Cui & Lee, 2020*)	301.1K	Social Posts User engagements News Articles	English	Binary (fake, real)	Distance Supervision
MM-COVID (*Li et al., 2020b*)	11.1K	Twitter	Multi-lingual	Binary (fake, real)	Manually
ArCOV19-Rumors (*Haouari et al., 2020b*)	9.4K	Twitter	Arabic	Three classes (false, true, other)	Manually
Alqurashi et al. (2021)	8.7K	Twitter	Arabic	Binary (misleading not-misleading)	Manually
COVID-19-FAKES (*Elhadad, Li & Gebali, 2020*)	0.4K	Twitter	Arabic + English	Binary (fake, real)	Automatically (13 ML algorithm)
Mahlous & Al-Laith (2021) (a)	2.5K	Twitter	Arabic	Binary (fake, genuine)	Manually
Mahlous & Al-Laith (2021) (b)	14.9K	Twitter	Arabic	Binary (fake, genuine)	Automatically
CLEF-2021 CheckThat! Lab (task 3A) (*Shahi, Struß & Mandl, 2021*)	1.2K	News articles	Multi-lingual	Four classes (false, true, partially false, other)	Manually
FakeCovid (*Shahi & Nandini, 2020*)	5.1K	Several Social media platforms	Multi-Lingual	Three classes (false, true, partially false)	Manually
Alsudias & Rayson (2020)	2K	Twitter	Arabic	Three classes (false, true, unrelated)	Manually
CMU-MisCOV19 (*Memon & Carley, 2020*)	0.5K	Twitter	English	Multi-class (17 classes)	Manually
Li et al. (2020a)	3K	Weibo	English	Multi-class (eight Situational classes)	Manually
ArCorona (*Mubarak & Hassan, 2020*)	8K	Twitter	Arabic	Multi-class (17 classes) (Mixing situational and misinformation classes)	Manually
AraCOVID19-MFH (*Ameur & Aliane, 2021*)	10.8K	Twitter	Arabic	Ten independent tasks (each with 2–4 classes)	Manually
Alam et al. (2020)	722	Twitter	Arabic	Ten independent tasks (each with 2–3 classes)	Manually
Out dataset (ArCOV19-MCM)	6.7K	Twitter	Arabic	Multi-class (19 misinformation classes)	Manually
Out dataset (ArCOV19-MLM)	6.7K	Twitter	Arabic	Multi-label (19 misinformation classes)	Manually
Out dataset (ArCOV19-Sit)	4.2K	Twitter	Arabic	Multi-class (Six situational classes)	Manually

explored methods in detecting COVID-19 misinformation with an accuracy of 86.2%. *Al-Yahya et al. (2021)* present a comparative study of various neural networks and transformer-based language models for detecting COVID-19 misinformation from Arabic text. The experimental results demonstrate that transformer-based models outperform the neural network-based solutions with a significant improvement in the F1 score from 83.0% by the best neural network-based model(GRU) to 95.0% by the best transformer-based

model (QARiB). *Kumari (2021)* submitted the top performing system to CLEF2021 CheckThat! (shared task 3A) for fake news detection with a macro F1-score of 88.0%. They employed a BERT-based model trained with the shared task's training data and an extensive amount of additional data obtained from various fact-checking websites.

DATASET

This section describes the steps we followed to collect and annotate our data.

Data collection

We decided to use previously fact-checked tweets to minimize the effort needed for veracity verification and focus on assigning fine-grained misinformation and situational information class labels. We acquired most of the tweets in our dataset from available annotated binary COVID-19 misinformation datasets collected in different periods of the pandemic. We avoided the datasets labeled automatically by a community of ML classifiers. In this research, we used the following datasets:

- *ArCOV19-Rumors* (*Haouari et al., 2020b*) is an Arabic misinformation dataset of tweets manually labeled into "false," "true," or "other," we only used the tweets under "false" and "true" classes.
- A misinformation Detection sample of 2,000 tweets was manually annotated into "true," "false," and "unrelated" (*Alsudias & Rayson, 2020*). This sample was originally taken randomly by the authors from a large general-purpose unlabeled COVID-19 dataset of six million unlabeled tweets they collected using COVID-19-related keywords.
- The Arabic COVID-19 binary misinformation dataset (*Alqurashi et al., 2021*) encompasses tweets manually classified into tweets with misinformation (false) and tweets with not-misleading information (true). The annotation is performed by relying on trusted resources such as WHO and the Ministry of Health in Saudi Arabia.

We decided to use multiple datasets collected and released during different periods of the pandemic instead of relying completely on a single dataset. Our goals is to collect data that covers as many COVID-19-related subjects as possible (*e.g.*, virus properties, symptoms, prevention methods, statistics, governmental measures, *etc.*). We believe that the topics that individuals and groups vocalize about the pandemic vary with time, so we tried to obtain as comprehensive data as possible. Several studies show noticeable topic trends that change over time and in response to major events (*Ordun, Purushotham & Raff, 2020*; *Bogdanowicz & Guan, 2022*).

Vaccine-related tweets. Since we depend on datasets annotated and released before any COVID-19 vaccines were developed and approved, we manually collected vaccine-related tweets and verified their veracity. Specifically, We manually collected 12 common and verified claims about the COVID-19 vaccines from the Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/) and Mayo Clinic (https://www.mayoclinic.org/). Table 2 presents these claims along with their English translation. After that, we extracted a set of keywords from the claims to retrieve potentiality-related tweets. We considered keyword alternatives and repeatedly modified the search. We used Twitter API for tweet

Table 2 Verified claims used to collect COVID-19 vaccine-related tweets and their English translations. The first six claims are confirmed to be true, and the rest are false.

1	COVID-19 vaccine is safe and effective.	لقاح كوفيد آمن وفعال.
2	There is no evidence that COVID-19 vaccines cause infertility and miscarriage.	لا يوجد دليل على أن لقاحات كوفيد تسبب العقم والإجهاض.
3	None of the authorized and recommended COVID-19 vaccines cause people to test positive on other viral tests.	لا يتسبب أي من لقاحات كوفيد المصرح بها والموصى بها في جعل اختبار الأشخاص إيجابيًا في الاختبارات الفيروسية
4	COVID-19 vaccines cause you to be magnetized.	تتسبب لقاحات كوفيد في أن تكون ممغنطًا.
5	COVID-19 vaccines alter DNA.	لقاحات كوفيد تغير الحمض النووي.
6	COVID-19 vaccines are injected as electronic implant chips to track and monitor people.	يتم حقن لقاحات كوفيد كرقاقات إلكترونية لتتبع الأشخاص ومراقبتهم.
7	Current COVID-19 vaccines do not protect against the COVID-19 variants.	لا تحمي لقاحات كوفيد الحالية من المتحورات.
8	There are severe side effects of COVID-19 vaccines.	هناك آثار جانبية خطيرة للقاحات كوفيد.
9	The natural immunity I get from being sick with COVID-19 is better than the immunity I get from COVID-19 vaccination.	إن المناعة الطبيعية التي أحصل عليها من الإصابة بفيروس كوفيد أفضل من المناعة التي أحصل عليها من لقاح كوفيد.
10	COVID-19 vaccines are not safe because the were rapidly developed and tested.	لقاحات كوفيد ليست آمنة لأنه تم تطويرها واختبارها بسرعة.
11	The ingredients in COVID-19 vaccines are dangerous.	المكونات الموجودة في لقاحات كوفيد خطيرة.
12	I won't need to wear a mask after I get vaccinated for COVID-19.	لن أحتاج إلى ارتداء قناع بعد أخذ اللقاح.

retrieval. After removing the irrelevant, we got 795 tweets left. We assigned "true" to the tweet that mostly paraphrases or restates a true claim or opposes a false claim (Figs. 1A and 1B). Likewise, the tweet that mostly matches a false claim or counters a true claim is labeled as a "false" tweet (Figs. 1C and 1D).

Our initial collection consists of the tweets with "true" and "false" labels from previously collected datasets after removing the duplicated tweets besides the vaccine-related tweets we've collected manually.

Data annotation

We started by defining our misinformation and situational information annotation guidelines. We initially took a sample of the 1,000 tweets and inspected them to check the misinformation and situational information types that frequently appear. We also benefited from annotation schemes previously designed for multi-class misinformation detection (*Memon & Carley, 2020*) and situational information identification (*Li et al., 2020a*) in the English text. We have reached our fine annotation schemes after going through several class refinement iterations during which some classes have been merged, split and even deleted.

Most existing fake news detection datasets contain only two or three main misinformation classes, such as "false/true" and "fake/real/unrelated." In this work, we define a total of 19 fine-grained classes of misinformation, such as "fake cure," "true cure,"

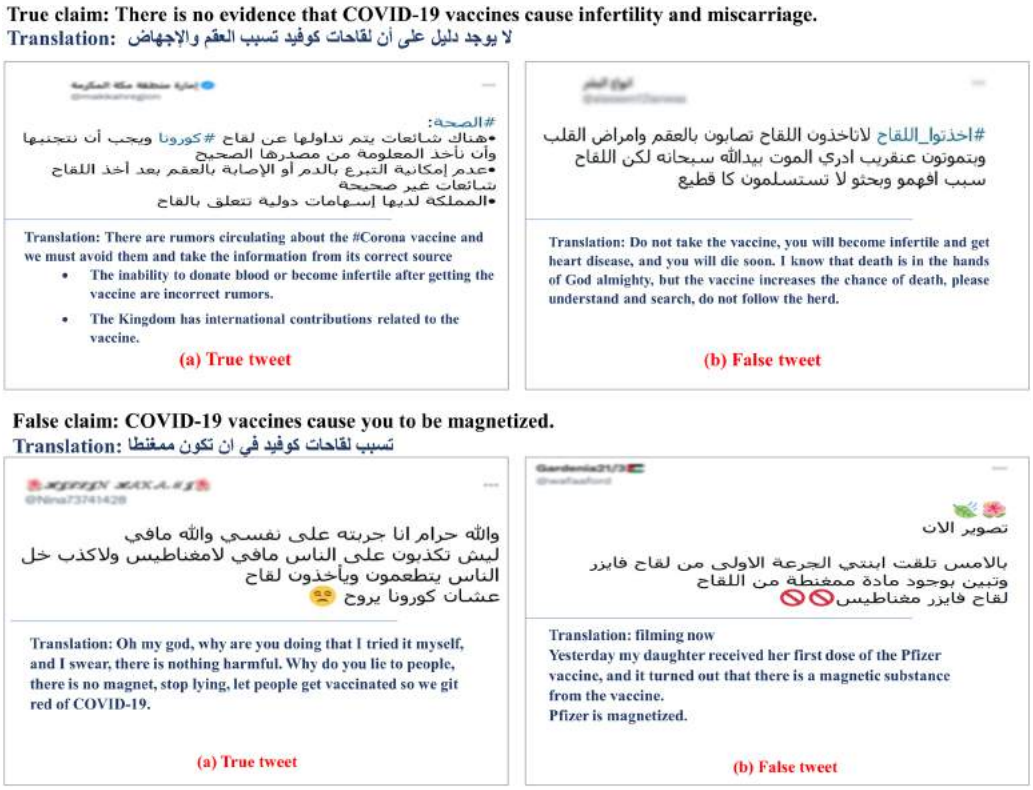


Figure 1 Tweets in (A) and (B) are "true" and "false" tweets collected for a true claim, respectively. (C) and (D) are "true" and "false" tweets collected for a false claim, respectively (with their translations).
Full-size DOI: 10.7717/peerj-cs.1151/fig-1

"fake prevention methods," "true prevention methods," "conspiracy theory," *etc*. We believe that the high-level binary (or ternary) classification does not consider the varying degrees of tweets' fact-checking worthiness or the harm-level misleading tweets could bring to society. Some types of misleading information about COVID-19 are worth more urgent fact-checking than others. Identifying the fact-checking-worthiness of Twitter data helps decide which claims should receive a higher priority from humans to perform further manual fact-checking (*Nakov et al., 2021*; *Barrón-Cedeño et al., 2020*). For instance, tweets that promote a fake cure for COVID-19 should get more attention from authorities and social media platforms and be removed urgently, whereas it could be sufficient to flag tweets with less sinister content with warning labels. Since we depend on tweets that are already fact-checked into "false" or "true," we only focus on assigning the fine-grained topic to each tweet based on the tweet content according to the definition provided in Table 3. For example, the class "true symptom" is assigned that tweets originally classified under the class "true," and at the same time, mention one or more COVID-19 symptoms. However, we had to verify the truth of some ambiguous tweets, which are not clear to which aspect of the tweet the truthiness/falseness has been assigned, as we describe shortly in the annotation challenges. "False Symptom" was one of the classes initially defined in our annotation scheme, but we excluded it since we didn't find any tweets belonging to it.

Table 3 Description and distribution of the misinformation classes used in ArCOV19-MCM and ArCOV19-MLM.

Class	Description	Count in MCM	Count in MLM
True virus origin and properties	Verified information about the origin and characteristics of the virus (*e.g.*, people of all ages could be infected by COVID-19).	221	221
False virus origin and properties	Incorrect or unverified information about the origin and characteristics of the virus (*e.g.*, COVID-19 is killed by hot climate).	282	311
True spread method	Verified information about how the virus spreads between people (*e.g.*, close contact).	116	137
False spread method	Incorrect or unverified information about how the virus spreads among people (house flies and mosquito bites transmit COVID-19).	242	269
True prevention method	A verified method to prevent or mitigate the chance of being infected by COVID-19. Examples include wearing a mask, social distancing, and washing hands.	535	813
False prevention method	An incorrect or unverified method to prevent or mitigate the chance of being infected by COVID-19 (*e.g.*, dust prevents the spread of COVID-19).	85	91
True vaccine info	Correct information about COVID-19 vaccines and their effectiveness that has been verified by a competent authority such as FDA (*e.g.*, COVID-19 vaccines provide strong protection against serious illness, hospitalization and death).	278	278
False vaccine info	Incorrect or unverified information about COVID-19 vaccines and their effectiveness (*e.g.*, the COVID-19 vaccine causes infertility).	442	443
True treatment or cure	A method of treatment to reduce the symptoms or ease the pain in COVID-19 patients, such as getting plenty of sleep and liquids. This class also includes any treatment or cure method approved by competent authorities.	68	69
Fake treatment or cure	Any treatment or cure that is not verified by a competent authority such as WHO, FDA, or CDC or any treatment or cure proved harmful (*e.g.*, hydroxychloroquine and bleach).	761	783
True symptom	A symptom that has appeared in COVID-19 patients (*e.g.*, Fever, Dry cough, and Fatigue).	112	180
True governmental measure	It is a measure that has already been taken by a government or an official authority, such as closing country borders, shops and worship places.	397	397
False governmental measure	A measure that any government or official authority has not taken.	120	120
Conspiracy theory	A conspiracy story such as the "virus is a bioweapon" and "Bill Gates is planning to microchip the world through COVID-19 vaccines."	312	320
Hate and exclusion	Offensive speeches or opinions that carry feelings of hate, racism, xenophobia, or blaming others caused by the epidemic (*e.g.*, "China virus").	37	38
Causing anxiety and fear	Stories that cause anxiety and spread panic among people.	176	229
Marketplace rumors	Rumors about products, brands, and companies related to the pandemic that could be harmful or confuse people and society.	24	24
Statistics	Statistics about COVID-19, including casualties, deaths, people recovered *etc*.	370	409
Irrelevant	The tweet cannot be classified under any of the classes above.	2,104	2,104
Total		6,682	7,235

Tweets that belong to "true vaccine info" and "false vaccine info" received their fine-grained misinformation classes during the collection process, as explained in the previous section. We found that 8.29% of tweets belong to more than one class, *e.g.*, a tweet

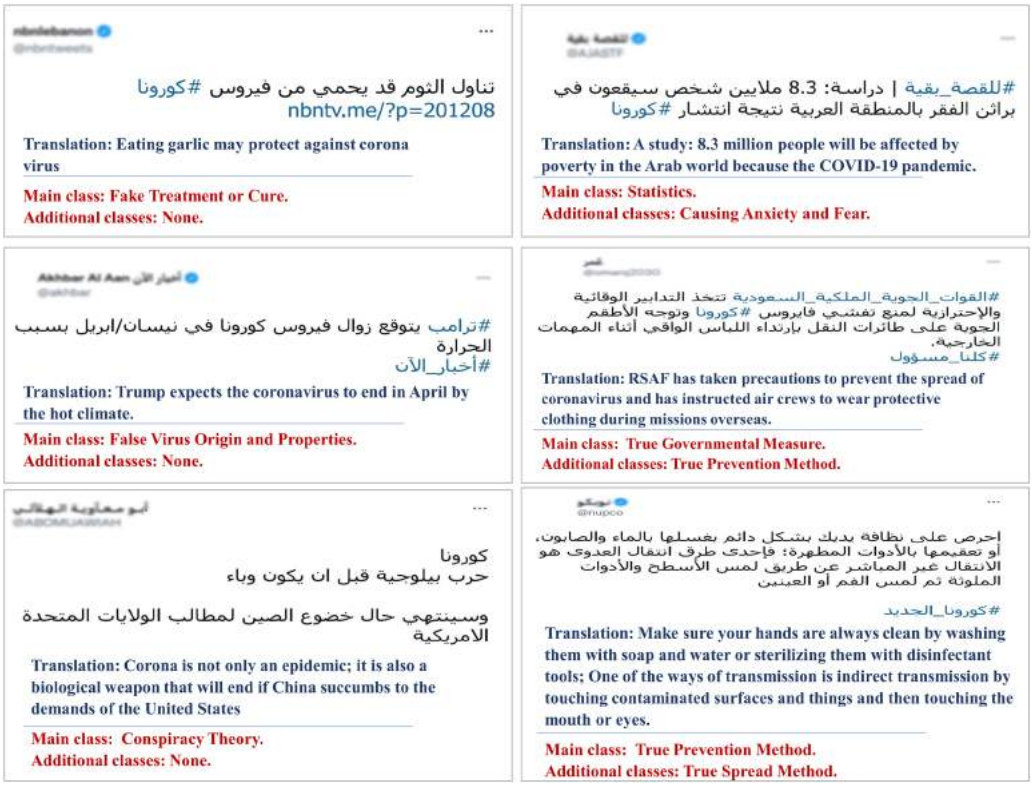


Figure 2 **Examples of tweets with multiple class labels, the examples to the left have only one primary class, while those to the right are assigned multiple labels.**
Full-size DOI: 10.7717/peerj-cs.1151/fig-2

promotes a "conspiracy theory" that causes anxiety and fear (*i.e.*, belongs to the type "causing anxiety and fear"). Thus we found it necessary to allow multiple labels. Accordingly, we assigned each tweet the "primary" class to which the main point in the tweet belongs and then assigned all other applicable classes as additional class labels. As a result, we define two versions of our **Ar**abic **COV**ID-**19**-related tweet dataset (ArCOV19) with:

- **M**ulti-**C**lass **M**isinformation annotations (ArCOV19-MCM), in which each tweet is only assigned a primary class.
- **M**ulti-**L**abel **M**isinformation annotations (ArCOV19-MLM) in which each tweet might have received additional classes besides the primary class.

Figure 2 presents example tweets from our dataset with their primary and additional misinformation class labels.

Situational information classes. Situational information helps individuals and authorities to understand the situation during emergencies (*Li et al., 2020a*). Examples include actionable information such as asking for immediate help, offering support, advice and cautions, *etc.* Identifying situational information is important for the concerned authorities to sense the public's mood, fill information gaps with the public, and develop proper emergency response strategies (*Yan & Pedraza-Martinez, 2019*). Since situational

Table 4 The definitions and the distribution of classes in ArCOV19-Sit.

Class	Description	Count
Caution and advice	Precautions to face the pandemic, such as washing hands, staying at home, wearing masks, avoiding travel, and encouraging compliance with precautionary measures against COVID-19 employed by authorities.	1,222
Notifications or measures	Outbreak announcements, such as the number of infections, recovery cases and deaths, and the measures taken by the concerned authorities, such as closing country borders, shops, schools and worship places and banning or restricting events.	612
Help seeking/offering	Seeking or offering support in the form of materials, money, goods, or service or asking for emotional help because of depression, fears *etc*.	28
Doubt casting and criticizing	Questioning local government officials, the Red Cross, and other related initiatives for inaction or blaming some of the public to mislead others.	594
Refute rumors	Responses against circulating rumors.	94
Non-situational	Information that can not be classified under any of the classes above.	1,682
Total		4,232

information aims to help individuals and authorities conceive the situation during major events (*e.g.*, crisis), we excluded the tweets that are originally identified as false tweets (in the original datasets) from our dataset version that is annotated with situational classes, as we believe that it is illogical relying on data with misinformation to understanding the situation during major events. We define six types of situational information and assign one situational information type to each tweet to build (ArCOV19-Sit), a version of **ArCOV19** with multi-class **Sit**uational annotations. These classes, along with their definitions and counts, are presented in Table 4.

Annotation challenges. We faced the following challenges during the annotation process.

1. We initially started with two classes related to the COVID-19 statistics in our tweet collection, "true statistics" and "false statistics." However, we found that statistics change drastically over time, so we combined them into one class called "statistics."
2. As aforementioned, we use previously fact-checked tweets to minimize the effort needed to check their veracity. However, By looking at tweets containing information about the disease, such as cures, treatments, symptoms, vaccines, *etc*., we found that some of these tweets also tell other types of stories about other entities involved, such as the person who promoted this fake cure or warned from taking that vaccine. We found that the original binary misinformation class is sometimes assigned based on whether that "other story" happened or not instead of assigning the misinformation class based on the falseness or trueness of the scientific information in the tweet. For example, by looking at the tweet presented in Fig. 3A, Donald Trump's proclaimed a prescription to treat COVID-19 patients by injecting them with antiseptics. The binary class assigned to this tweet is "True" because Trump did suggest that, but the fine-grained misinformation class we gave is "false treatment or cure." This forced us to not depend on the assigned binary class and manually verify the misinformation of all multi-story tweets that involve scientific information about the virus from credible resources, including WHO, CDC, Mayo Clinic and the FDA. We added a verification link (a link to the page we used



Figure 3 **Examples of tweets that posed additional challenges during data annotation.**
Full-size DOI: 10.7717/peerj-cs.1151/fig-3

for verification) to each of these tweets in our data. We removed any "multi-story" tweet that we could not verify the truthfulness of from our dataset.

3. At the beginning of the pandemic, some drugs, such as hydroxychloroquine that is used to treat malaria, were approved by FDA for emergency use to treat COVID-19. Later, FDA revoked this emergency use authorization (*Office of the Commissioner, 2022*). Since we depend on Twitter datasets that have been released relatively at the early stages of the pandemics, we couldn't rely on the original binary class label to classify the cure-and-treatment tweets into "false treatment or cure" and "true treatment or cure" as it might have changed as with tweet in Fig. 3B. To this end, we manually performed this classification of all of the cure-and-treatment tweets by referring to the credible resources mentioned earlier.

Annotation quality. We randomly selected 1.5K tweets along with their multi-label and multi-class misinformation classes (*i.e.*, from ArCOV19-MCM and ArCOV19-MLM). Then, we recruited a second annotator, who has no access to the labels by the first annotator, to perform a second round of annotation. The inter-annotator agreement of the misinformation annotations in ArCOV19-MCM is 97.83% using Cohen's kappa coefficient. The inter-annotator agreement of the misinformation annotations in ArCOV19-MLM is 95.78, using Krippendorff's alpha reliability coefficient (*Bhowmick, Basu & Mitra, 2008*). As for situational information classes, we excluded the tweets that didn't receive situational classes from the 1,500 examples mentioned above (*i.e.*, the tweets that were originally classified as false tweets in the binary misinformation datasets from which we collect our data). The inter-annotator agreement of the situational annotations in ArCOV19-Sit is 96.79 using Cohen's kappa coefficient.

The high annotation agreements of the three dataset versions prove the annotations' quality, given that some classes are more semantically related than others. The main reason for annotations divergence in ArCOV19-MLM is the subjectivity in assigning one of the classes of the applicable classes to the tweets that belong to more than one class, which assures the importance of dealing with misinformation classification as a multi-label

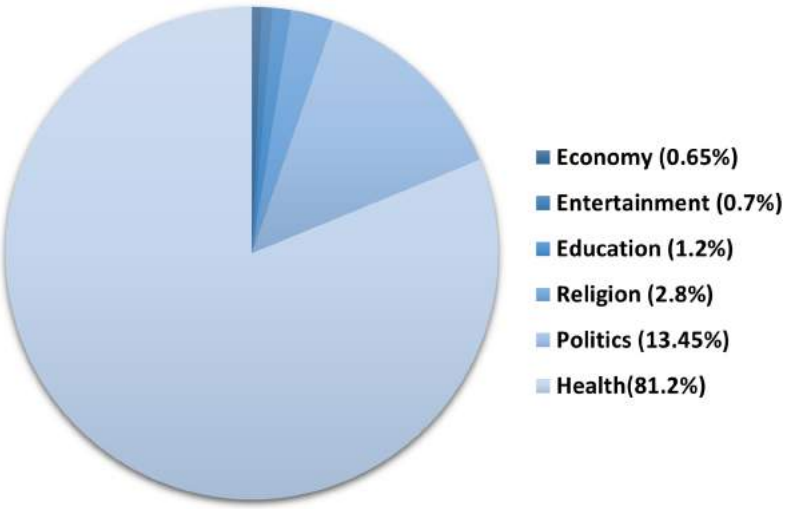


Figure 4 **Topics distribution in ArCOV19-MCM.** Full-size DOI: 10.7717/peerj-cs.1151/fig-4

classification problem. The testing sets of ArCOV19-MCM, ArCOV19-MLM, and ArCOV19-Sit are sampled from the tweets that received full agreement between the two annotators to guarantee high-quality testing data.

Data analysis and discussion

In this section, we performed an exploratory analysis of our dataset. The versions ArCOV19-MCM and ArCOV19-MLM contain 6,681 tweets annotated with multi-class and multi-labeled misinformation classes. ArCOV19-Sit has 4,236 tweets. The label cardinality of ArCOV19-MLM is 1.08. Label cardinality is the average number of labels per tweet (*Nam et al., 2014*). It's an indication of how hard the classification problem is. The low cardinality denotes a less-complex label space (*Sorower, 2010*). Our dataset has considerably low cardinality, which suggests that the classification task is not very difficult.

Topic analysis. The global outbreak of COVID-19 affected almost all aspects of life. To assess the impact of COVID-19 on the different sectors, we annotated the 2,000 tweets from ArCOV19-MCM with the general topic they belong to, including health, education, economy, politics, religion and entertainment. Figure 4 shows that health is the most frequent topic, followed by politics. We believe that the entertainment sector received less attention from social media users not because it has not been affected but because it has become a less-vital matter compared to health.

Class distribution. Figures 5A and 5B represent the class distributions of the training and the testing sets in ArCOV19-MCM, respectively. Both sets have imbalanced class distributions. The class sizes are consistent between the testing and the training sets. The most common class in the training set is the class "irrelevant," with 1,884 tweets, followed by "fake treatment or cure," with 639 tweets. The least populated class is "true treatment or cure." A plausible explanation is is that COVID-19 is relatively new; hence there is limited evidence regarding specific antivirals that may work against it (*Singh & de Wit, 2022*). The

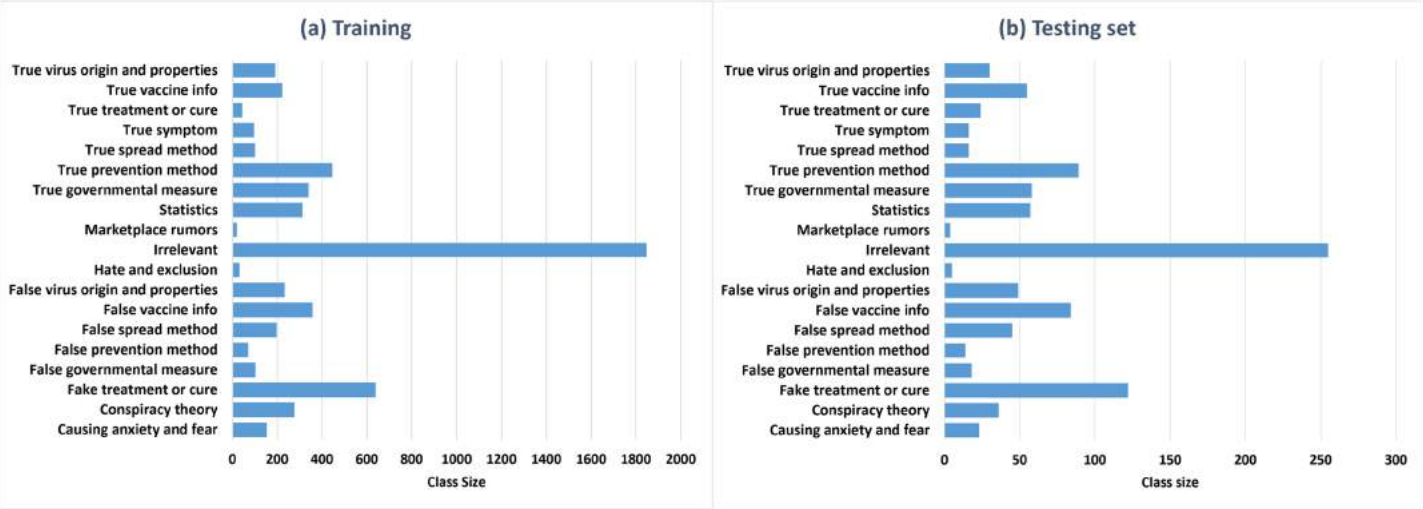


Figure 5 **The class distributions of the training (A) and the testing (B) sets of ArCOV19-MCM.** Full-size DOI: 10.7717/peerj-cs.1151/fig-5

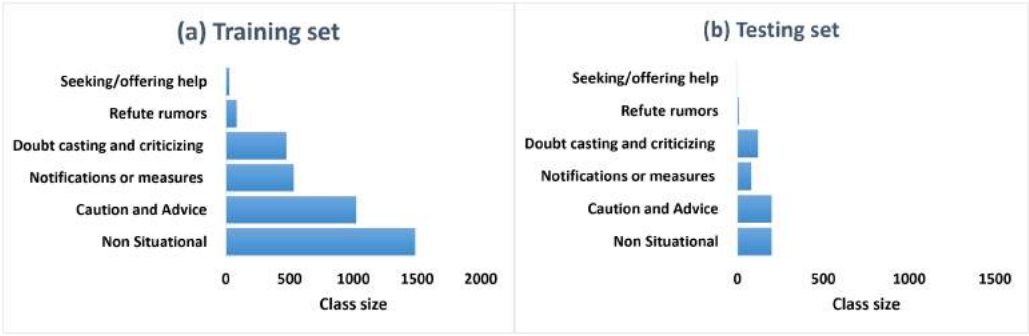


Figure 6 **The class distributions of the training (A) and the testing (B) sets of ArCOV19-Sit.** Full-size DOI: 10.7717/peerj-cs.1151/fig-6

class distributions of the training and the testing sets of ArCOV19-Sit are shown in Fig. 6. The distributions are heavily skewed. The smallest class is "Help seeking/offering," and the "non-situational" class is the highest populated. The training set's most commonly populated situational information class is "caution and advice," followed by "notification and measures".

Tweet length. The length of the tweets in the dataset is an important property as it influences the dimensionality of the text's feature vector that represents the input to the ML model. For example, TF-IDF is usually too sparse for short text. As shown in Fig. 7, our tweets have a length range from 3 to 87 tokens, most of them with 20 to 50 tokens. We also observed no variations in tweet lengths among different classes.

Word cloud. Figure 8 provides the word clouds for the six misinformation classes along with their translations by picking the top 200 words in each class. Several interesting insights can be observed from these word clouds. For instance, some of the frequent words that appear in the class "true vaccine info" are آمن, الصحة and تقديم that are

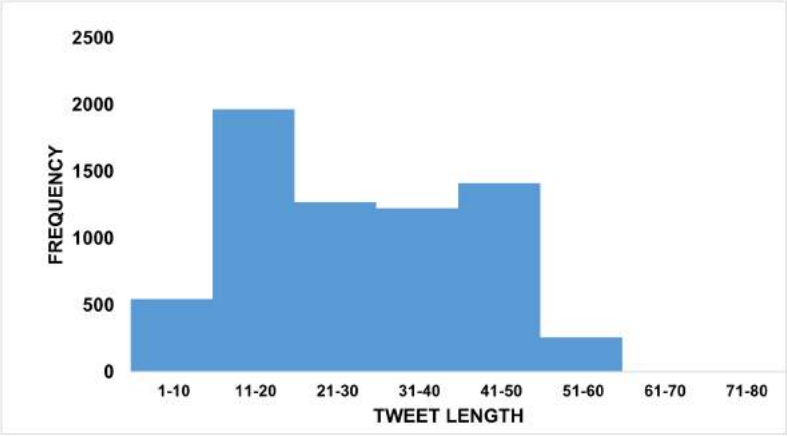


Figure 7 **Tweet length distributions in ArCOV19-MCM and ArCOV19-MLM.**
Full-size DOI: 10.7717/peerj-cs.1151/fig-7

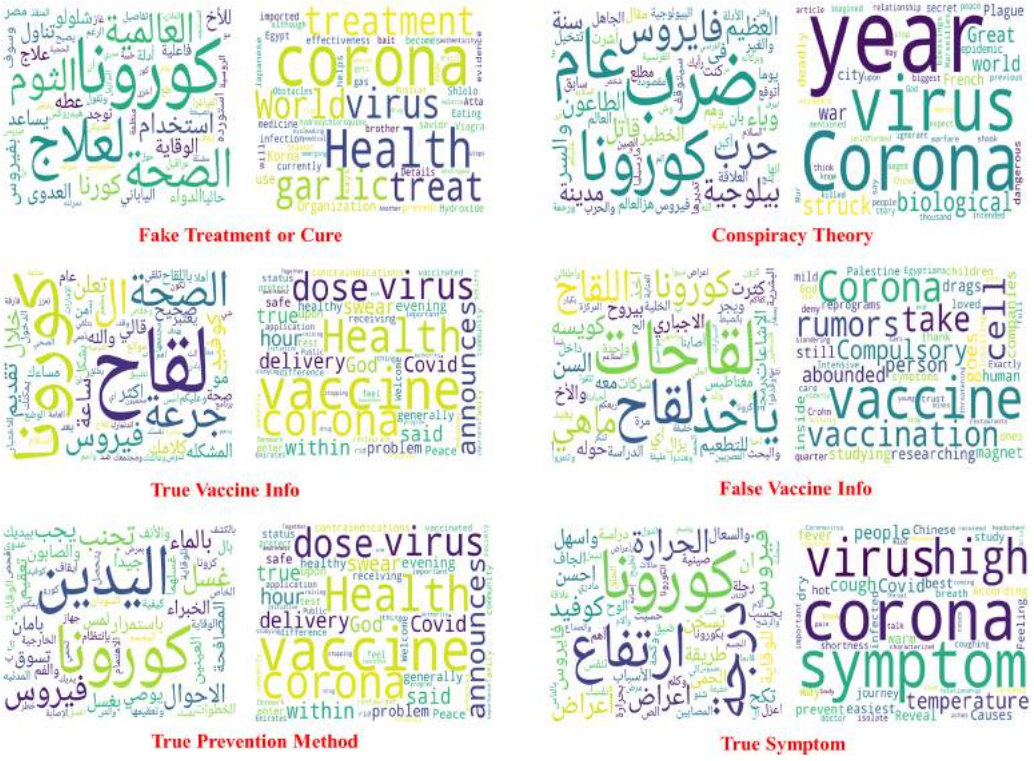


Figure 8 **Word clouds of six of the misinformation classes along with their translated clouds.**
Full-size DOI: 10.7717/peerj-cs.1151/fig-8

translated to *safe*, *health*, *vaccine* and *delivery*. These positive words encourage vaccine intake. However, by looking at the frequent words of the class "False vaccine info" such as مغناطيس, اشاعات, اجباري and that are translated into *magnet*, *rumors*, and *compulsory*. These words increase vaccine hesitancy. Likewise, the tweets under the class "true prevention methods" have a frequent appearance of words that encourage personal

hygiene, such as غسل and اليدين (translated to *wash* and *hand*). In the class "conspiracy theory", the most frequently used words are بيولوجية and حرب (translated to *biological* and *war*), promoting one of the most widespread COVID-19-related conspiracy theories since the start of the pandemic (*Matt Burgess, 2021*).

BASELINES AND EVALUATION

Classification methods

This section explores the performance of several Ml and DL methods to perform multi-class misinformation classification, multi-label misinformation classification and multi-class situational information classification, representing baselines for future research using our data.

Multi-class classification

We performed classification into the misinformation and the situational classes by training several ML and transformer-based classifiers using ArCOV19-MCM and ArCOV19-Sit. The following is an overview of the classification models used in our experiments. We describe the features we utilize to transform the raw tweets into feature vectors representing the ML classifiers' input.

Features. In this study, we use three popular features to convert all tweets to numerical vectors before being fed to the ML algorithms. (1) Term Frequency/Inverse Document Frequency (TF-IDF) is a numerical statistic measure computed by multiplying two metrics to assign a weight of importance to each term in the tweet. First, Term Frequency (TF) measures the importance of a term in a tweet (document). Second, IDF (t) is a weight that indicates how frequently a term appears in the dataset. The terms that are common in the dataset receive low IDF scores. (2) N-Grams are contiguous sequences of N elements in a given text. An element in an N-gram could be a character, word, or even a sentence and is decided based on the task. It is obtained by sliding a window of size N over the text. An N-gram of size one is referred to as a unigram; 2-gram is called a bigram, and 3-gram is known as a trigram. We experimented with character-level bigrams. (3) Stylistic features are a type of features that help to discover different authoring and writing styles for the authors of fake news *vs* authors of legitimate news (*Hussein, 2019*). We extracted 14 hand-crafted stylistic features, including the number of words, characters, hashtags, user mentions (@), URLs, links, emojis, ellipses, digits, commas, period, punctuation marks, exclamation marks, percentage marks and question marks.

Machine learning algorithms. We employ three well-known ML algorithms to perform multi-class misinformation classification: SVM (*Pisner & Schnyer, 2020*), random forest (RF) (*Cutler, Cutler & Stevens, 2012*), and XGBoost (*Chen & Guestrin, 2016*), as they have proven to be successful in various Arabic text classification tasks including sentiment analysis (*Abooraig et al., 2018*), and hate speech detection (*Aljarah et al., 2020*).

Transformer-based models. Transformers (*Vaswani et al., 2017*) are neural network architecture that applies the mechanism of self-attention to selectively weigh the significance of each part of the input text. They have been pretrained on huge unsupervised general corpora. Pretrained transformer-based models such as BERT (*Devlin et al., 2018*)

and post-BERT models like RoBERTa (*Liu et al., 2019*) and ALBERT (*Lan et al., 2019*) have achieved state-of-the-art performance for several NLP tasks classification and generation tasks. In this work, we use five transformer models:

- **AraBERT** (https://huggingface.co/aubmindlab/bert-base-arabertv02) is a BERT model pretrained on 200 million Arabic MSA sentences gathered from various resources based on Google's BERT-Base configuration (*Antoun, Baly & Hajj, 2020*).
- **AraBERT-COV19** (https://huggingface.co/moha/arabertc19) is a version of AraBERT pretrained using an Arabic COVID-19 Twitter dataset (*Ameur & Aliane, 2021*). The dataset consists of 1.5 million raw multi-dialect Arabic tweets about the COVID-19 pandemic.
- **AraBERTv02-Twitter** (https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter) is a transformer-based model obtained by continuing the pretraining AraBERT using 60M multi-dialect Arabic tweets.
- **MARBERTv02** (https://huggingface.co/UBC-NLP/MARBERTv2) MARBERT is a transformer-based model that exploits large-to-massive scale multi-dialect datasets. It has been evaluated on several NLP tasks, including sentiment analysis, Question Answering and dialect identification. MARBERTv02 is MARBERT that has been further fine-tuned using additional data with longer sequences to handle the task that needs to capture longer contexts (*Abdul-Mageed, Elmadany & Nagoudi, 2021*).
- **XLM-R** (https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base) is a transformer-based masked language model trained jointly on one hundred languages, including Arabic, to improve the cross-lingual understanding tasks (*Barbieri, Espinosa-Anke & Camacho-Collados, 2022*; *Conneau et al., 2020*).

Multi-label classification

Unlike multi-class classification, in which each instance is associated with an exactly single class label, multi-label classification is a supervised learning problem in which the data instance can simultaneously belong to more than one class. Multi-label classification methods can be roughly divided into problem transformation and algorithm adaptation. The problem transformation method involves transforming a multi-label problem into one or more independent binary or multi-class classification tasks. The final multi-label prediction for a new instance is determined by aggregating the classification results from the individual classifiers. On the other hand, algorithm adaptation methods extend traditional learning algorithms to handle multiple labels directly (*Tsoumakas & Katakis, 2007*). We use two problem transformation methods: binary relevance (BR) (*Boutell et al., 2004*) and label powerset (LP) (*Tsoumakas, Katakis & Vlahavas, 2009*). LinearSVC and RF are adopted as base classifiers.

Experimental setup

Evaluation metric and model selection

Following prior works, We evaluate the performance of our baseline models by reporting macro-averaged and micro-averaged Precision (P), Recall (R) and F1 scores in addition to

Table 5 The optimal hyperparameters of the ML algorithms used in multi-class misinformation and situational information classification.

Classifier	Hyper-parameters	Misinformation classification	Situational classification
SVM (linear kernel)	C	1	1
	Gamma	1	1
RF (entropy criterion)	n_estimators	50	200
	max_depth	100	500
	max_features	sqrt	sqrt
XGBoost	max_depth	10	10
	n_estimators	100	100
	random_state	42	100
	eta	0.3	0.3
	subsample	1	1

Accuracy. We use the micro-averaged F1 score as the primary metric for comparison. In addition, we employ the Jaccard score as an additional evaluation metric of multi-label classification. To evaluate the performance of the considered multi-class and multi-label misinformation classification models, we sampled 1,000 tweets out of the data that the annotators agreed upon as testing sets of dataset versions ArCOV19-MCM and ArCOV19-MLM. The goal is to maintain the same division across all models to allow a fair comparison. The testing set of ArCOV19-Sit consists of 608 tweets that resulted from removing the tweets that were originally annotated as false (in the source binary datasets) from the 1,000-tweet sample. As we mentioned earlier, some tweets in textbfArCOV19-MLM received additional classes besides the primary class. It is worth noting that we don't distinguish the primary class label from the additional class labels during modeling.

Hyper-parameter tuning

This section presents hyper-parameter tuning performed to the learning techniques considered in this study. We applied the grid search (*Liashchynskyi & Liashchynskyi, 2019*) to identify the optimal hyper-parameters for the ML algorithms and tried selected values to tune the transformer-based classifiers. We use the training set of each version to fit the models. The hyper-parameters are tuned using 10-fold cross-validation. The models that achieve the best cross-validation micro-averaged F-scores are applied to the testing set to report the final results.

Table 5 presents the optimal hyper-parameters of SVM, RF, and XGboost used in the multi-class misinformation and situational information classification tasks. We use the SVM linear kernel (LinearSVC) with the One-*vs*-One approach implementation provided by scikit-learn (https://scikit-learn.org/stable/). Regarding the BERT models, including AraBERT, mBERT and AraBERT COV19, the AdamW optimizer is used with a learning rate of 4e−5, the batch size is set to 8, and the maximum sequence length is 128. We used early stopping to tune the number of epochs and found that three epochs yielded the best performance. We adopted the default hyperparameters of LinearSVC and RF used for multi-label misinformation classification.

Table 6 Performance of the multi-class misinformation classification models trained and evaluated using ArCOV19-MCM. Boldface represents the performance scores of the best ML classifier and the best transformer-based for each metric.

Classifier	Accuracy	Macro-P	Macro-R	Macro-F1	Micro-F1
SVM	**77.3%**	80.55%	**66.86%**	**71.38%**	**77.3%**
RF	56.1%	**85.68%**	41.74%	48.55%	56.1%
XGBoost	72.39%	79.01%	61.78%	67.19%	72.39%
AraBERT	80.3%	83.46%	73.79%	76.24%	80.3%
AraBERT-COV19	**81.6%**	80.69%	**77.96%**	**78.76%**	**81.6%**
AraBERTv02-Twitter	81.1%	**81.27%**	75.95%	77.02%	81.1%
MARBERTv2	80.1%	76.39%	72.98%	73.15%	80.1%
XLM-R	79.1%	75.32%	71.16%	71.21%	79.1%

Results and discussion

This work defines three classification tasks: multi-class misinformation classification, multi-label misinformation classification, and multi-class situational information classification. Several Ml classifiers and transformers-baseline models are trained and tested for each classification task representing baselines for future research on our data. This section reports the experimental results of these classifiers for each classification task and discusses reported performance.

Multi-class misinformation classification

We applied the ML classifiers, XGBoost, SVM and RF, and five transformer-based pretrained models: AraBERT, AraBERT-COV19, AraBERTv02-Twitter, MARBERTv2 and XLM-R to perform multi-class misinformation classification. All models are trained using the training set of ArCOV19-MCM and evaluated using the corresponding testing set. The ML classifier was trained using a combination of n-grams and TF-IDF stylistic features. We extracted the word-level unigrams and bigrams and char-level bigrams and trigrams. Then, we computed the TF-IDF of each type. The final vector representing each tweet is a concatenation of these TF-IDF vectors and the stylistic features after being normalized to have the same scale as the TF-IDF scores. These vectors represent the input of the ML models. The values of accuracy, macro-averaged and micro-averaged precision, recall, and F1 scores are repaired. All the experimental results of this task are summarized in Table 6. It is worth mentioning that the accuracy and the micro-averaged F1 scores for multi-class classification problems are mathematically equivalent; thus, they have identical values. The best results are highlighted in bold font.

Results of the ML classifiers. Several observations and insights can be drawn from these results. First, among the compared ML algorithms, SVM achieved the best accuracy of 77.3%, macro-averaged recall of 66.86% and macro-averaged F1 score of 71.38%, followed by XGBoost. Second, the micro-averaged F1 scores are consistently higher than the macro-average F1 scores of the three models. This is because the dataset ArCOV19-MCM has imbalanced class distribution. For example, 31.6% of the instances belong to the class "irrelevant," and the rest are distributed unevenly over the remaining classes (See Fig. 5).

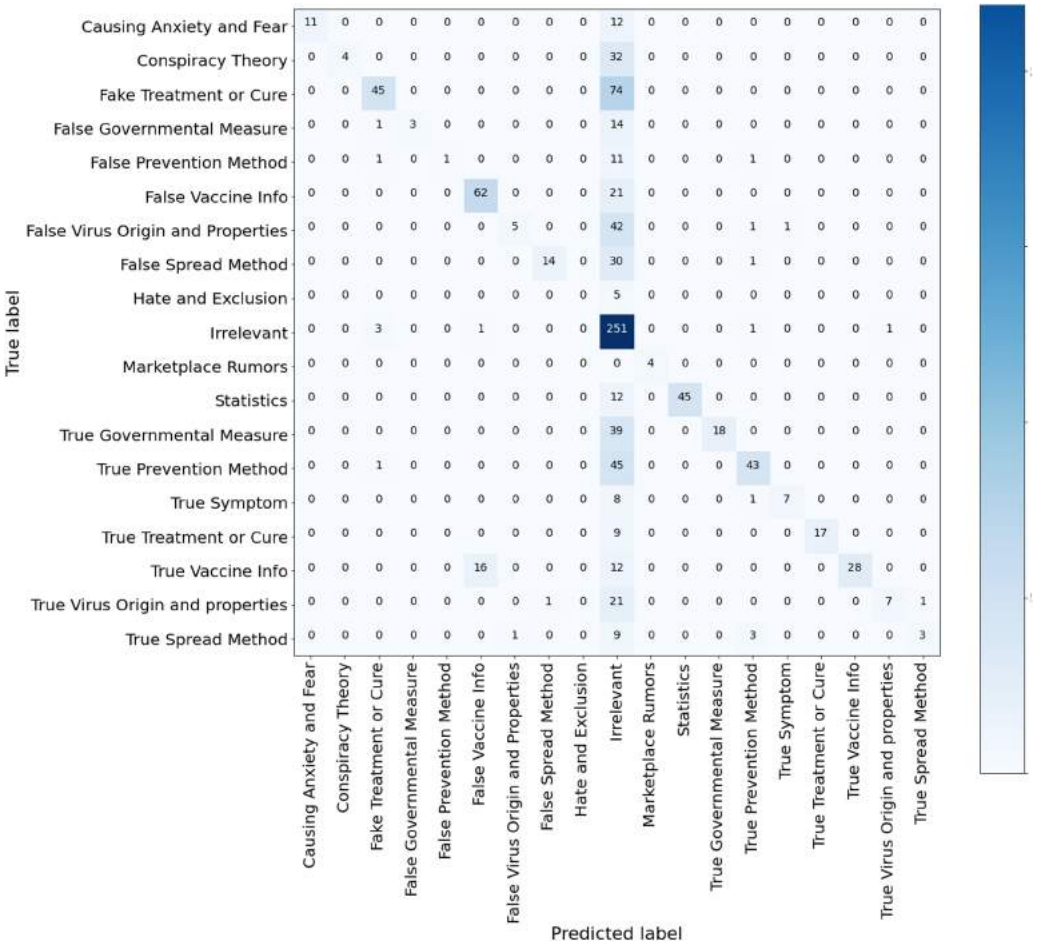


Figure 9 **Confusion matrix RF model trained for multi-class misinformation classification.**
Full-size DOI: 10.7717/peerj-cs.1151/fig-9

We thought in the beginning that it would be better if we excluded the class "irrelevant" from the data used to establish the baselines. Then, we decided to keep it and see which model is more efficient in dealing with this challenging yet common problem. Third, RF has the worst recall of 41.74% and the highest precision compared to all classifiers, including the Bert models, probably because RF is very sensitive to imbalanced data. Since the RF algorithm samples a subset from the training set to train every decision tree, these subsets will contain fewer tweets from the minority classes. Thus, RF could fail to recognize these tweets amplifying the skewed data effect (*Chen, Liaw & Breiman, 2004*). By observing the confusion matrix of the RF classifiers depicted in Fig. 9, most of the tweets in the minority classes were categorized incorrectly under the class "irrelevant." The same reason also explains the RF's high precision score, as most of the tweets are classified under the majority classes, irrelevant, and the percentage of the predicted majority-classes predictions that are correctly classified is high. Although SVM and XGboost significantly outperform RF, they also suffer from the uneven distribution of the classes, manifested by the differences between the macro-averaged and the micro-averaged F1 scores of these two

classifiers. These findings suggest that there is room for improvement by employing techniques that deal with imbalanced class distributions (*He & Garcia, 2009*; *Fernández et al., 2018*), or by collecting more data to correct the variations in class sizes. This is particularly important to attain the main purpose of defining fine-grained misinformation classes. Another observation can be obtained from RF's Confusion matrix in Fig. 9 that the model mixed up the tweets that belong to the classes "True vaccine info" and "False vaccine info," probably because the tweets in these two classes use similar vocabulary (as shown in Fig. 8), limiting the ability of the ML models to differentiate them. A possible solution is by augmenting our data with features extracted from Author profiles and user engagements instead of depending solely on the tweet's contexts.

Results of the transformer-based models. Table 6 reveals that AraBERT-COV19 followed by AraBERTv02-Twitter outperforms all models, including the ML methods, with an accuracy of 83.20%, macro-precision of 81.82%, macro-recall of 78.71% and a macro-F1 score of 79.52%. We think that AraBERT-COV19 surpassed AraBERT since it is tuned using 1.5M multi-dialect Arabic tweets related to COVID-19, making it familiar with the Arabic COVID-19 multi-dialect vocabulary. Also, since AraBERT-COV19 is trained on a Twitter dataset, it is more suitable to represent tweets in which text can be misspelled, abbreviated and has irregular syntax. In contrast, the AraBERT was mostly pretrained in Modern Standard Arabic (MSA), thus less capable of dealing with multi-dialectal Arabic than AraBERT COV19. Following AraBERT-COV19, AraBERTv02-Twitter also performs comparably well since it is a version of AraBERT obtained by continuing the pretraining AraBERT using 60M multi-dialect Arabic tweets. By observing the confusion matrix of AraBERT-COV19 in Fig. 10, AraBERT-COV19 is remarkably more capable of dealing with the imbalanced data and capturing the semantic differences between the defined fine-grained misinformation classes, confirming the efficiency of the transformers and their ability to transfer knowledge between domains and proving the importance of pretraining the BERT-based models on COVID-19-related unsupervised data.

The least-performing transformer-based method is XLM-R. A plausible explanation is that XLM-R is pretrained using multi-lingual data and usually outperformed by the monolingual models pretrained with large language-specific datasets and rich vocabularies (*Virtanen et al., 2019*; *Alammary, 2022*). MARBERT achieved comparable performance to AraBERT in the micro-averaged F1-score but suffered from a performance gap in terms of the macro-averaged F-scores, although MARBERT is pretrained with larger data. A systematic review of BERT Models for various Arabic text classification problems (*Alammary, 2022*) shows that AraBERT outperformed MARBERT in several tasks and *vice versa*. It also shows that a large pretraining *corpus* does not necessarily guarantee better performance. It would be important to study different BERT-based models by connecting their performance with the distribution of the classes in the considered tasks.

Multi-class situational information classification

Table 7 presents the experimental results of multi-class situational classification models, trained and evaluated using the ArCOV19-Sit dataset. We employ the same baseline

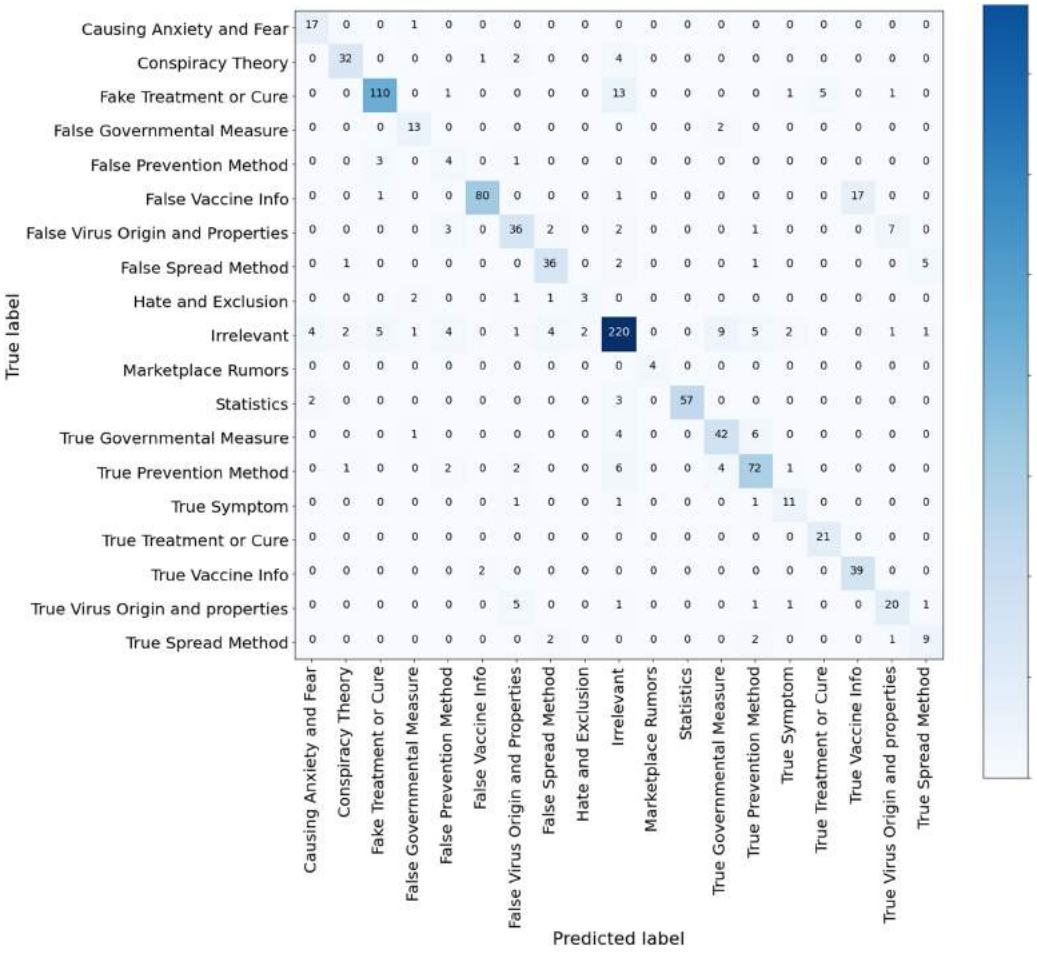


Figure 10 Confusion matrix AraBERT COV19 model trained for multi-class misinformation classification. Full-size DOI: 10.7717/peerj-cs.1151/fig-10

Table 7 Performance of the multi-class situational information classification models trained and evaluated using ArCOV19-Sit. Boldface represents the performance scores of the best ML classifier and the best transformer-based for each metric.

Classifier	Accuracy	Macro-P	Macro-R	Macro-F1	Micro-F1
SVM	**73.39%**	54.38%	52.79%	**53.07%**	**73.39%**
RF	66.83%	**59.16%**	44.18%	46.78%	66.83%
XGBoost	73.23%	56.54%	49.71%	51.47%	73.23%
AraBERT	77.5%	67.56%	58.79%	61.24%	77.5%
AraBERT-COV19	78.81%	**73.96%**	**77.3%**	**73.45%**	78.81%
AraBERTv02-Twitter	77.99%	64.66%	76.62%	66.58%	77.99%
MARBERT	**79.31%**	58.9%	58.11%	**58.28%**	**79.31%**
XLM-R	77.83%	61.19%	58.57%	**59.37%**	77.83%

models, features and evaluation measures used in the multi-class misinformation classification tasks. These results of ML algorithms are in concordance with the results presented in the previous section. Among the ML classifiers, SVM achieves superior

Table 8 Performance of the multi-label misinformation classification models trained and evaluated using ArCOV19-MLM. Boldface represents the best performance scores for each metric.

Classifier	Binary Relevance (BR)			Label Powerset(LP)		
	Jaccard	Macro-F1	Micro-F1	Jaccard	Macro-F1	Micro-F1
RF	39.35%	48.61%	55.25%	62.65%	55.4%	62.54%
Linear SVC	60.29%	67.84%	75.23%	**76.3%**	**68.89%**	**76.69%**

performance to XGBoost and RF with an accuracy of 73.39%, macro recall of 52.79%, and macro-F1 score of 53.07%. Comparing various pretrained models, AraBERT-COV19 outperformed all others in macro-averaged precision, recall, and F1 score, while MARBERT achieved the best micro-averaged F1 score. Based on the results of this section and the previous section, we hypothesize that MARBERT is more sensitive to the imbalanced class distribution, emphasizing the need to study the relationship between different BERT pretrained models and the label space. Once again, XLM-R performed the worst in most of the cases.

Multi-label misinformation classification

This section presents the experimental result of the BR and LP multi-label classification algorithm trained using the ArCOV19-MLM dataset by utilizing LinearSVC and RF as base classifiers. The macro-averaged and micro-averaged F1 scores and Jaccard scores are reported in Table 8. Notably, the linear SVC outperformed RF as a base classifier in BR and LP, and this is consistent with our previous findings that explain why RF is more sensitive to the skewed distribution of the data. Table 8 also shows that LP outperforms BR using all measures regardless of the base classifier. BR decomposes the multi-label learning task into 19 independent binary learning tasks, one class label. Each binary classifier outputs "true" if the tweet belongs to that class or "false" if it doesn't. Then binary classifier votes individually to obtain the final predictions. In other words, the predicted label set consists of the class labels that the corresponding binary classifiers output "true." On the other hand, the LP approach transforms a multi-label problem into a single-label multi-class problem and applies a multi-class classification approach to solve the problem. The classes include all class combinations that appeared in the training set. By observing the class distribution of ArCOV19-MLM presented in Table 3, we can see that both types of transformation performed by LP and BR increase the class imbalance. However, the approach that is used by BR transformation causes a more severe data imbalance not only for the minor classes but also for the moderately-populated classes. This suggests applying adaptation-based multi-label classification approaches which don't include data transformation could improve the performance for this task, which we leave for future work.

Discussion of potential applications

In light of the experimental results that confirm the practical utility of our data, several parties can potentially benefit from the presented work, including, but not limited to, social media platforms, health professionals, public authorities and decision-makers. Assigning a

specific misinformation sub-type to COVID-19-related social media posts can be employed by the social media platforms to direct the information to the most interested audience. It also helps assess the degree of harm misinformation can bring people and take the appropriate measures accordingly. Possible measures could range from assigning warning labels to the susceptible post to removing posts with malicious content or even suspending the account that shares the misleading posts entirely (*Levush, 2019*; *Alam et al., 2020*). Health professionals can utilize the suggested models to identify the health-related misinformation among the existing diverse misinformation themes, which is considered the first step in fighting their proliferation. Fine-grained misinformation and situational information can be exploited by the public authorities and decision-makers to understand the situation and the public mood during the pandemic peak, assisting in making better decisions to guide people, support them and save their lives.

CONCLUSION

The outbreak of COVID-19 has accompanied the massive spread of misleading information about the disease through social media. Combatting COVID-19 misinformation on online platforms is crucial to the success of official authority efforts to control the pandemic. Manual fact-checking organizations may not be able to keep up with the misleading information that appears daily during the pandemic, creating a need for computational models capable of automatically identifying misinformation. Therefore, several research efforts have been dedicated to creating labeled misinformation datasets and utilizing them to develop and evaluate various ML and DL models. Most of these datasets have binary annotation in which the text (*e.g.*, tweet, post, *etc.*) is labeled as credible or non-credible. However, the misleading news stories related to COVID-19 are not equal, and some types of misleading information (*e.g.*, fake cure) could be life-threatening and should be instantly removed from social media platforms. For this reason, we've collected COVID-19-related Arabic tweet data containing 6,681 tweets. We manually annotated our data with fine-grained multi-label misinformation classes. Furthermore, we annotated a subset of this dataset (after removing the misleading tweets) with multi-class situational information classes. Identifying the situational information is very important for authorities and individuals to assess the situation during crises and understand the public's moods and needs. We performed several types of analysis to prove the quality of our annotations. Besides, we defined several classification tasks and applied multiple ML and transformer-based classifiers to train misinformation and situational information classification models. The experimental results confirm the validity of our data in developing different types of ML and DL models. The results also prove that AraBERT-COV19, a transformer-based model pretrained on large unsupervised COVID-19 Twitter data, outperformed the other methods by a significant margin with a micro-F1 score of 81.6% and a macro-F1 score of 73.4% for the multi-class misinformation and situational classification tasks, respectively. Regarding the multi-label misinformation task, the label power set using linear SVC as a base classifier outperformed the other tested methods with micro-F1 score of 76.69%. For future work, we could consider incorporating additional features related to author profiling and user engagement

to improve ML classifiers' performance instead of relying only on features extracted from the tweet contents. Another possible direction is to enrich our dataset versions and flatten their skewed class distribution by collecting more data, which greatly supports deploying the trained systems using this data in real-life applications.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Rasha Obeidat conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, data annotation quality analysis, and approved the final draft.
- Maram Gharaibeh conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, data annotation quality analysis, and approved the final draft.
- Malak Abdullah conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Yara Alharahsheh performed the experiments, analyzed the data, prepared figures and/or tables, data annotation quality analysis, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data and code are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1151#supplemental-information.

REFERENCES

Abdul-Mageed M, Elmadany A, Nagoudi EMB. 2021. ARBERT & MARBERT: deep bidirectional transformers for Arabic. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7088–7105.

Abdul-Mageed M, Elmadany A, Nagoudi EMB, Pabbi D, Verma K, Lin R. 2020. Mega-COV: a billion-scale dataset of 100+ languages for COVID-19. *ArXiv preprint*. DOI 10.48550/arXiv.2005.06012.

Abdullah M, Altiti O, Obiedat R. 2022. Detecting propaganda techniques in english news articles using pre-trained transformers. In: *2022 13th International Conference on Information and Communication Systems (ICICS)*. Piscataway: IEEE, 301–308.

Abooraig R, Al-Zu'bi S, Kanan T, Hawashin B, Al Ayoub M, Hmeidi I. 2018. Automatic categorization of Arabic articles based on their political orientation. *Digital Investigation* **25(5)**:24–41 DOI 10.1016/j.diin.2018.04.003.

Al-Yahya M, Al-Khalifa H, Al-Baity H, AlSaeed D, Essam A. 2021. Arabic fake news detection: comparative study of neural networks and transformer-based approaches. *Complexity* **2021**:1–10 DOI 10.1155/2021/5516945.

Alam F, Shaar S, Nikolov A, Mubarak H, Martino GDS, Abdelali A, Dalvi F, Durrani N, Sajjad H, Darwish K, Al-Homaid A, Zaghouani W, Caselli T, Danoe G, Stolk F, Bruntink B, Nakov P. 2020. Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *ArXiv preprint* DOI 10.48550/arXiv.2005.00033.

Alammary AS. 2022. BERT models for Arabic text classification: a systematic review. *Applied Sciences* **12(11)**:5720 DOI 10.3390/app12115720.

Aljarah I, Habib M, Hijazi N, Faris H, Qaddoura R, Hammo B, Abushariah M, Alfawareh M. 2020. Intelligent detection of hate speech in Arabic social network: a machine learning approach. *Journal of Information Science* **47(4)**:165551520917651 DOI 10.1177/0165551520917651.

Alomari E, Katib I, Albeshri A, Mehmood R. 2021. COVID-19: detecting government pandemic measures and public concerns from Twitter Arabic data using distributed machine learning. *International Journal of Environmental Research and Public Health* **18(1)**:282 DOI 10.3390/ijerph18010282.

Alonso MA, Vilares D, Gómez-Rodríguez C, Vilares J. 2021. Sentiment analysis for fake news detection. *Electronics* **10(11)**:1348 DOI 10.3390/electronics10111348.

Alqurashi S, Hamoui B, Alashaikh A, Alhindi A, Alanazi E. 2021. Eating garlic prevents COVID-19 infection: detecting misinformation on the Arabic content of Twitter. *ArXiv preprint*. DOI 10.48550/arXiv.2101.05626.

Alsudias L, Rayson P. 2020. COVID-19 and Arabic Twitter: how can Arab world governments and public health organizations learn from social media? In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

AlTarrah D, AlShami E, AlHamad N, AlBesher F, Devarajan S. 2021. The impact of coronavirus COVID-19 pandemic on food purchasing, eating behavior, and perception of food safety in Kuwait. *Sustainability* **13(16)**:8987 DOI 10.3390/su13168987.

Ameur MSH, Aliane H. 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news and hate speech detection dataset. *ArXiv preprint*. DOI 10.48550/arXiv.2105.03143.

Antoun W, Baly F, Hajj H. 2020. AraBERT: transformer-based model for Arabic language understanding. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France, 9–15.

Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, Artemova E, Tutubalina E, Chowell G. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia* **2(3)**:315–324 DOI 10.3390/epidemiologia2030024.

Barbieri F, Espinosa-Anke L, Camacho-Collados J. 2022. XLM-T: multilingual language models in Twitter for sentiment analysis and beyond. In: *Proceedings of the LREC*. Marseille, France, 20–25.

Barrón-Cedeño A, Elsayed T, Nakov P, Da San Martino G, Hasanain M, Suwaileh R, Haouari F, Babulkov N, Hamdan B, Nikolov A, Shaar S, Ali ZS. 2020. Overview of CheckThat! 2020: automatic identification and verification of claims in social media. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Cham: Springer, 215–236.

Barua Z, Barua S, Aktar S, Kabir N, Li M. 2020. Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science* **8**:100119 DOI 10.1016/j.pdisas.2020.100119.

Bhowmick PK, Basu A, Mitra P. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*. Manchester, UK, 58–65.

Bogdanowicz A, Guan C. 2022. Dynamic topic modeling of Twitter data during the COVID-19 pandemic. *PLOS ONE* **17(5)**:e0268669 DOI 10.1371/journal.pone.0268669.

Boutell MR, Luo J, Shen X, Brown CM. 2004. Learning multi-label scene classification. *Pattern Recognition* **37(9)**:1757–1771 DOI 10.1016/j.patcog.2004.03.009.

Chen T, Guestrin C. 2016. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*. 785–794.

Chen E, Lerman K, Ferrara E. 2020. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance* **6(2)**:e19273 DOI 10.2196/19273.

Chen C, Liaw A, Breiman L. 2004. A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform. *University of California, Berkeley* **110(1–12)**:24.

Ciampaglia GL, Mantzarlis A, Maus G, Menczer F. 2018. Research challenges of digital misinformation: toward a trustworthy web. *AI Magazine* **39(1)**:65–74 DOI 10.1609/aimag.v39i1.2783.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

Cui L, Lee D. 2020. COAID: COVID-19 healthcare misinformation dataset. *ArXiv preprint*. DOI 10.48550/arXiv.2006.00885.

Cutler A, Cutler DR, Stevens JR. 2012. Random forests. In: *Ensemble Machine Learning*. Berlin: Springer, 157–175.

Da San Martino G, Yu S, Barrón-Cedeno A, Petrov R, Nakov P. 2019. Fine-grained analysis of propaganda in news article. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 5636–5646.

De Sarkar S, Yang F, Mukherjee A. 2018. Attending sentences to detect satirical fake news. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, 3371–3380.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint*. DOI 10.48550/arXiv.1810.04805.

Dizikes P. 2018. Study: on Twitter, false news travels faster than true stories. *Available at https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308* (accessed 20 April 2018).

Elhadad MK, Li KF, Gebali F. 2020. COVID-19-FAKES: a Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. In: *International Conference on Intelligent Networking and Collaborative Systems*. Cham: Springer, 256–268.

Essam BA, Abdo MS. 2021. How do Arab tweeters perceive the COVID-19 pandemic? *Journal of Psycholinguistic Research* **50(3)**:507–521 DOI 10.1007/s10936-020-09715-6.

Farha IA, Magdy W. 2019. Mazajak: an online Arabic sentiment analyser. In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Florence, Italy, 192–198.

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. 2018. *Learning from imbalanced data sets.* Vol. 10. Cham: Springer.

Greene CM, Murphy G. 2021. Quantifying the effects of fake news on behavior: evidence from a study of COVID-19 misinformation. *Journal of Experimental Psychology: Applied* **27(4)**:773–784 DOI 10.1037/xap0000371.

Gupta RK, Vishwanath A, Yang Y. 2020. Global reactions to COVID-19 on Twitter: a labelled dataset with latent topic, sentiment and emotion attributes. *ArXiv preprint*. DOI 10.48550/arXiv.2007.06954.

Haouari F, Hasanain M, Suwaileh R, Elsayed T. 2020a. ArCOV-19: the first Arabic COVID-19 Twitter dataset with propagation networks. *ArXiv preprint*. DOI 10.48550/arXiv.2004.05861.

Haouari F, Hasanain M, Suwaileh R, Elsayed T. 2020b. ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection. *ArXiv preprint*. DOI 10.48550/arXiv.2010.08768.

He H, Garcia EA. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21(9)**:1263–1284 DOI 10.1109/TKDE.2008.239.

Helmstetter S, Paulheim H. 2021. Collecting a large scale dataset for classifying fake news tweets using weak supervision. *Future Internet* **13(5)**:114 DOI 10.3390/fi13050114.

Hossain TT. 2021. COVIDLies: detecting COVID-19 misinformation on social media. PhD thesis, UC Irvine.

Hossain T, Logan RL IV, Matsubara Y, Young S, Singh S. 2020. COVIDLies: detecting COVID-19 misinformation on social media. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Hussein A. 2019. Content-social based features for fake news detection model from Twitter. *International Journal of Advanced Trends in Computer Science and Engineering* **8(6)**:2806–2810 DOI 10.30534/ijatcse/2019/20862019.

Islam MS, Sarkar T, Khan SH, Kamal A-HM, Hasan SM, Kabir A, Yeasmin D, Islam MA, Chowdhury KIA, Anwar KS, Chughtai AA, Seale H. 2020. COVID-19-related infodemic and its impact on public health: a global social media analysis. *The American Journal of Tropical Medicine and Hygiene* **103(4)**:1621–1629 DOI 10.4269/ajtmh.20-0812.

Kabha R, Kamel A, Elbahi M, Hafiz A, Dafri W. 2020. Impact of fake news and myths related to COVID-19. *Journal of Content Community and Communication* **12**:270–279 DOI 10.31620/JCCC.12.20/25.

Khalifa H, Al-Absy MSM, Badran SA, Alkadash TM, Almaamari QA, Nagi M. 2020. COVID-19 pandemic and diffusion of fake news through social media in the Arab world. *Arab Media & Society* **30**:.

Kim HK, Ahn J, Atkinson L, Kahlor LA. 2020. Effects of COVID-19 misinformation on information seeking, avoidance, and processing: a multicountry comparative study. *Science Communication* **42(5)**:586–615 DOI 10.1177/1075547020959670.

Kleinberg B, van der Vegt I, Mozes M. 2020. Measuring emotions in the COVID-19 real world worry dataset. *ArXiv preprint*. DOI 10.48550/arXiv.2004.04225.

Kolluri NL, Murthy D. 2021. CoVerifi: a COVID-19 news verification system. *Online Social Networks and Media* **22(7)**:100123 DOI 10.1016/j.osnem.2021.100123.

Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Akl E, Baddour K. 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* **12(3)**:e7255 DOI 10.7759/cureus.7255.

Kumari S. 2021. NoFake at CheckThat! 2021: fake news detection using BERT. *ArXiv preprint*. DOI 10.48550/arXiv.2108.05419.

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. 2019. ALBERT: a lite BERT for self-supervised learning of language representations. *ArXiv preprint* DOI 10.48550/arXiv.1909.11942.

Levush R. 2019. Government responses to disinformation on social media platforms: Argentina, Australia, Canada, China, Denmark, Egypt, European Union, France, Germany, India, Israel, Mexico, Russian federation, Sweden, United Arab Emirates, United Kingdom. *Available at https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1180&context=scholcom*.

Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao T-L, Duan W, Tsoi KK-F, Wang F-Y. 2020a. Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on weibo. *IEEE Transactions on Computational Social Systems* **7(2)**:556–562 DOI 10.1109/TCSS.2020.2980007.

Li Y, Jiang B, Shu K, Liu H. 2020b. MM-COVID: a multilingual and multimodal data repository for combating COVID-19 disinformation. *ArXiv preprint*. DOI 10.48550/arXiv.2011.04088.

Liashchynskyi P, Liashchynskyi P. 2019. Grid search, random search, genetic algorithm: a big comparison for nas. *ArXiv preprint*. DOI 10.48550/arXiv.1912.06059.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. *ArXiv preprint* DOI 10.48550/arXiv.1907.11692.

Liu Y, Wu Y-F. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence* **32(1)**:11268 DOI 10.1609/aaai.v32i1.11268.

Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* **5(3)**:337–348 DOI 10.1038/s41562-021-01056-1.

Mahlous AR, Al-Laith A. 2021. Fake news detection in Arabic tweets during the COVID-19 pandemic. *International Journal of Advanced Computer Science and Applications* **12(6)**:120691 DOI 10.14569/IJACSA.2021.0120691.

Matt Burgess WU. 2021. New vaccine conspiracy theories are going viral in Arabic. *Available at https://www.wired.com/story/new-vaccine-conspiracy-theories-going-viral-arabic/*.

Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU. 2020. An "infodemic": leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. *Open Forum Infectious Diseases* **7**:ofaa258 DOI 10.1093/ofid/ofaa258.

Memon SA, Carley KM. 2020. Characterizing COVID-19 misinformation communities using a novel Twitter dataset. *ArXiv preprint*. DOI 10.48550/arXiv.2008.00791.

Mubarak H, Hassan S. 2020. ArCorona: Analyzing Arabic tweets in the early days of coronavirus (COVID-19) pandemic. *ArXiv preprint*. DOI 10.48550/arXiv.2012.01462.

Nakov P, Da San Martino G, Elsayed T, Barrón-Cedeno A, Míguez R, Shaar S, Alam F, Haouari F, Hasanain M, Babulkov N, Nikolov A, Shahi GK, Struß JM, Mandl T. 2021. The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: *European Conference on Information Retrieval*. Cham: Springer, 639–649.

Nam J, Kim J, Loza Mencía E, Gurevych I, Fürnkranz J. 2014. Large-scale multi-label text classification—revisiting neural networks. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer, 437–452.

Office of the Commissioner. 2022. Covid-19 frequently asked questions. *U.S. Food and Drug Administration*. *Available at https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-frequently-asked-questions*.

Ordun C, Purushotham S, Raff E. 2020. Exploratory analysis of COVID-19 tweets using topic modeling, UMAP, and DiGraphs. *ArXiv preprint*. DOI 10.48550/arXiv.2005.03082.

Oshikawa R, Qian J, Wang WY. 2020. A survey on natural language processing for fake news detection. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 6086–6093.

Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, Ekbal A, Das A, Chakraborty T. 2021. Fighting an infodemic: COVID-19 fake news dataset. In: *International Workshop on Combating On line Hostile Posts in Regional Languages during Emergency Situation*. Berlin: Springer, 21– 29.

Pisner DA, Schnyer DM. 2020. Support vector machine. In: *Machine Learning*. Amsterdam: Elsevier, 101–121.

Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. 2017. Automatic detection of fake news. *ArXiv preprint*. DOI 10.48550/arXiv.1708.07104.

Qazi U, Imran M, Ofli F. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* **12(1)**:6–15 DOI 10.1145/3404111.3404114.

Raza S, Ding C. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics* **13(4)**:335–362 DOI 10.1007/s41060-021-00302-z.

Shahi GK, Nandini D. 2020. Fakecovid-a multilingual cross-domain fact check news dataset for COVID-19. *ArXiv preprint*. DOI 10.48550/arXiv.2006.11343.

Shahi GK, Struß JM, Mandl T. 2021. Overview of the CLEF-2021 CheckThat! Lab: task 3 on fake news detection. In: *CLEF (Working Notes)*. 406–423.

Simione L, Vagni M, Gnagnarella C, Bersani G, Pajardi D. 2021. Mistrust and beliefs in conspiracy theories differently mediate the effects of psychological factors on propensity for COVID-19 vaccine. *Frontiers in Psychology* **12**:1657 DOI 10.3389/fpsyg.2021.683684.

Singh M, de Wit E. 2022. Antiviral agents for the treatment of COVID-19: progress and challenges. *Cell Reports Medicine* **3(3)**:100549 DOI 10.1016/j.xcrm.2022.100549.

Sorower MS. 2010. *A literature survey on algorithms for multi-label learning.* Vol. 18. Corvallis: Oregon State University, 1–25.

Su Z, McDonnell D, Wen J, Kozak M, Abbas J, Šegalo S, Li X, Ahmad J, Cheshmehzangi A, Cai Y, Yang L, Xiang Y-T. 2021. Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices. *Globalization and Health* **17(1)**:1–8 DOI 10.1186/s12992-020-00654-4.

Tacchini E, Ballarin G, Della Vedova ML, Moret S, de Alfaro L. 2017. Some like it Hoax: automated fake news detection in social networks. In: *2nd Workshop on Data Science for Social Good, SoGood 2017*. CEUR-WS, 1–15.

Tsoumakas G, Katakis I. 2007. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining* **3(3)**:1–13 DOI 10.4018/jdwm.2007070101.

Tsoumakas G, Katakis I, Vlahavas I. 2009. Mining multi-label data. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer, 667–685.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 5998–6008.

Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, Ginter F, Pyysalo S. 2019. Multilingual is not enough: BERT for Finnish. *ArXiv preprint* DOI 10.48550/arXiv.1912.07076.

Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, Eide D, Funk K, Katsis Y, Kinney R, Li Y, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade A, Wang K, Wang NXR, Wilhelm C, Xie B, Raymond D, Weld DS, Etzioni O, Kohlmeier S. 2020. CORT-19: The COVID-19 open research dataset. *ArXiv preprint* DOI 10.48550/arXiv.2004.10706.

Wani A, Joshi I, Khandve S, Wagh V, Joshi R. 2021. Evaluating deep learning approaches for COVID-19 fake news detection. In: *International Workshop on Combating online hostile Posts in Regional Languages during Emergency Situation*. Berlin: Springer, 153–163.

Yan L, Pedraza-Martinez AJ. 2019. Social media for disaster management: operational value of the social conversation. *Production and Operations Management* **28(10)**:2514–2532 DOI 10.1111/poms.13064.

Yang Q, Alamro H, Albaradei S, Salhi A, Lv X, Ma C, Alshehri M, Jaber I, Tifratene F, Wang W, Gojobori T, Duarte CM, Gao X, Zhang X. 2020. SenWave: monitoring the global sentiments under the COVID-19 pandemic. *ArXiv preprint* DOI 10.48550/arXiv.2006.10842.

Zhou X, Mulay A, Ferrara E, Zafarani R. 2020. ReCOVery: a multimodal repository for COVID-19 news credibility research. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3205–3212.

Zhou X, Zafarani R. 2020. A survey of fake news: fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* **53(5)**:1–40 DOI 10.1145/3395046.

PeerJ Computer Science

Addressing religious hate online: from taxonomy creation to automated detection

Alan Ramponi[1], Benedetta Testa[2], Sara Tonelli[1] and Elisabetta Jezek[2]

[1] Fondazione Bruno Kessler, Trento, Italy
[2] Dipartimento di Studi Umanistici, Università di Pavia, Pavia, Italy

ABSTRACT

Abusive language in online social media is a pervasive and harmful phenomenon which calls for automatic computational approaches to be successfully contained. Previous studies have introduced corpora and natural language processing approaches for specific kinds of online abuse, mainly focusing on misogyny and racism. A current underexplored area in this context is religious hate, for which efforts in data and methods to date have been rather scattered. This is exacerbated by different annotation schemes that available datasets use, which inevitably lead to poor repurposing of data in wider contexts. Furthermore, religious hate is very much dependent on country-specific factors, including the presence and visibility of religious minorities, societal issues, historical background, and current political decisions. Motivated by the lack of annotated data specifically tailoring religion and the poor interoperability of current datasets, in this article we propose a fine-grained labeling scheme for religious hate speech detection. Such scheme lies on a wider and highly-interoperable taxonomy of abusive language, and covers the three main monotheistic religions: Judaism, Christianity and Islam. Moreover, we introduce a Twitter dataset in two languages—English and Italian—that has been annotated following the proposed annotation scheme. We experiment with several classification algorithms on the annotated dataset, from traditional machine learning classifiers to recent transformer-based language models, assessing the difficulty of two tasks: abusive language detection and religious hate speech detection. Finally, we investigate the cross-lingual transferability of multilingual models on the tasks, shedding light on the viability of repurposing our dataset for religious hate speech detection on low-resource languages. We release the annotated data and publicly distribute the code for our classification experiments at https://github.com/dhfbk/religious-hate-speech.

Subjects Artificial Intelligence, Computational Linguistics, Data Mining and Machine Learning, Natural Language and Speech, Network Science and Online Social Networks
Keywords Natural language processing, Abusive language detection, Religious hate speech detection

Submitted 31 May 2022
Accepted 16 September 2022
Published 15 December 2022

Corresponding author
Alan Ramponi, alramponi@fbk.eu

Academic editor
Arkaitz Zubiaga

Additional Information and Declarations can be found on page 27

DOI **10.7717/peerj-cs.1128**

Copyright
2022 Ramponi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

The growing popularity of social networks has created opportunities for like-minded people to locate each other online and create communities sharing values, as well as social, political or religious views. This is particularly evident on Twitter, where

following/follower relations, and reply/quote/retweet interactions ease the creation of more or less connected communities of users (*Boyd, Golder & Lotan, 2010*; *Lee, Hwalbin & Okhyun, 2015*; *Garimella, Weber & Choudhury, 2016*). Despite the positive effects of novel social connections, however, this possibility to bypass traditional media gatekeepers, paired with the high accessibility and potential anonymity of the users, have greatly contributed to the proliferation of online hate (*Christopherson, 2007*; *Johnson et al., 2019*). Several types of targets have been identified and analyzed in past literature, from specific individuals such as political candidates (*Grimminger & Klinger, 2021*) or female videogame makers (*Gray, Buyukozturk & Hill, 2017*), to groups and categories such as immigrants (*Sánchez-Junquera et al., 2021*) and journalists (*Charitidis et al., 2020*).

Given the pervasiveness and magnitude of abusive language online, numerous computational approaches to tackle this problem have been proposed in the past within the NLP community (*Fortuna & Nunes, 2018*), one of the first being the seminal work in *Waseem & Hovy (2016)*, addressing sexist and racist tweets. Research on online abuse has since then grown to include specific workshops (*Waseem et al., 2017a*; *Kumar et al., 2020*; *Akiwowo et al., 2020*) and shared tasks (*Zampieri et al., 2019*; *Zampieri et al., 2020*), and to cover a large number of languages (*Carmona et al., 2018*; *Basile et al., 2019*; *Corazza et al., 2020*; *Ranasinghe & Zampieri, 2022*). Besides research on online abuse in general terms, some works have tried to categorize different hateful messages online, identifying a number of possible targets and subtargets (*Salminen et al., 2018b*). Among them, racism and gender-based hatred have been extensively studied, creating also target-specific resources and detection systems. Racism has been analyzed from different perspectives, ranging from detection (*Onabola et al., 2021*) to annotators' bias (*Sap et al., 2019*; *Larimore et al., 2021*). As regards gender-based hatred online, there have been specific tasks devoted to detecting abusive language towards women (*Fersini, Rosso & Anzovino, 2018*; *Fersini, Nozza & Rosso, 2020*), as well as works analyzing gendered stereotypes in automatic sentence completion (*Nozza, Bianchi & Hovy, 2021*). Misogyny has also been analyzed at a fine-grained level, proposing a novel taxonomy of misogynistic language occurrences (*Zeinert, Inie & Derczynski, 2021*). Gender issues have been the focus also of the few research activities around automatic detection of microaggressions (*Breitfeller et al., 2019*). By contrast, religious hate online is a rather understudied problem within the NLP community, despite being an important and impactful societal issue (*Albadi, Kurdi & Mishra, 2018*). Indeed, according to some studies religious hate can lead to individuals' radicalization when there is no exposure to information that would challenge these ideas and beliefs (*Thompson, 2011*; *Awan, 2017*). It is therefore important to have a better understanding of how religious hate is expressed online, which forms of offensive language are typically employed with this type of target, and whether there are differences across different religions. This would not only lead to a better knowledge of the phenomenon, but also it would allow the development of more accurate and effective systems to detect religious hate online. In this work we therefore focus on this target, considering the world's three major monotheistic religions: Judaism, Islam and Christianity. To manually and automatically identify religious hate in online conversations, we adapt the definition of Islamophobia provided in *Vidgen & Yasseri (2020)* to encompass all the targets of interest by

defining religious hate as *indiscriminate negative attitudes or emotions directed at a religion or its believers*. Thus, in this study, we consider religious hate as a specific form of abuse or offense against a religious target. Details on our categorization are provided in the "A taxonomy for religious hate" section.

Our main contributions are the following: *(i)* we propose a taxonomy for religious hate covering the three main monotheistic religions, which is aligned with other taxonomies for abusive language classification (*Vidgen et al., 2021*; *Zeinert, Inie & Derczynski, 2021*); *(ii)* we create a new dataset containing English and Italian tweets, annotated according to the above taxonomy, which we release for research purposes; and *(iii)* we conduct a set of monolingual and cross-lingual experiments, including zero-shot transfer, to detect abusive language and religious hate speech. Our work shows that religious hate speech detection is a challenging problem even for state-of-the-art pretrained language models (with best results obtained by fine-tuned RoBERTa-based language models (*Liu et al., 2019*; 65.69 F_1 for Italian, and 64.94 F_1 for English), and highlights the viability of cross-lingual transfer *via* multilingual language models for detecting religious hate speech on languages in which annotated data for the task is not available or easily obtainable (with best results achieved by a fine-tuned XLM-RoBERTa language model (*Conneau et al., 2020*): 58.53 F_1 for Italian to English, and 60.27 F_1 for English to Italian).

The article is organized as follows. After reviewing related work, we first describe our novel taxonomy to categorize religious hate in online messages, consistent with the taxonomy proposed in *Zeinert, Inie & Derczynski (2021)* for misogyny and with the categories identified in *Vidgen et al. (2021)*, to which we added "abusive humor". We then detail the creation of a Twitter dataset containing two subsets to study religious hate, one in English and one in Italian. We describe the data annotation process and provide related documentation in the form of data and artifact statements (*Bender & Friedman, 2018*; *Ramponi & Tonelli, 2022*). Finally, we present monolingual and cross-lingual classification experiments on Italian and English data subsets for two tasks: abusive language detection and religious hate speech identification, followed by a quantitative and qualitative analysis of religious hate speech forms across languages and religions, a discussion and our conclusions.

NOTE: This article contains examples of language which may be offensive to some readers.

RELATED WORK

Taxonomies of online abuse

Since online hate is a complex, multi-faceted phenomenon, the need for a categorization of types of abuse has been addressed in several works, also because it can guide annotators in labeling hate speech data more accurately. *Waseem et al. (2017b)* present a typology of abusive language, distinguishing between individuals and groups as target and between implicit and explicit hate. A three-layered annotation has been introduced for the OffensEval shared tasks (*Zampieri et al., 2020*), where participants are required to distinguish between offensive and not offensive tweets, targeted and untargeted offenses, and to classify the target type (*i.e.,* individual, group and other). Other taxonomies include more fine-grained categories of abuse. For example, the annotation scheme proposed

by *Palmer et al. (2020)* revolves around the offensiveness of a message, the presence of slurs, adjectival nominalization, and distancing. *Sanguinetti et al. (2018)*, instead, focus on hate messages against immigrants and annotate hate intensity, aggressiveness, offensiveness, irony and stereotypes. For misogyny, at least two taxonomies have been recently proposed. The first one, by *Anzovino, Fersini & Rosso (2018)*, included five types of misogynistic language, *i.e.,* discredit, stereotype and objectification, sexual harassment, dominance, and derailing. The second one, by *Zeinert, Inie & Derczynski (2021)*, modified the previous one by adding benevolent sexism and neo-sexism and removing derailing, as a consequence of discussions with annotators. *Vidgen et al. (2021)*, instead, propose a general-purpose taxonomy of abusive content, encompassing any type of target and offense type. Abuse, in particular, can be identity-directed, affiliation-directed and person-directed, and the first two categories include derogation, animosity, threatening language, glorification of hateful entities, and dehumanization as sub-categories. Our taxonomy has been designed so to be compliant with *Vidgen et al. (2021)*'s by further specifying the taxonomy branch related to identity-directed with a religious target. We enrich the subcategories with abusive humor, while merging derogation and dehumanization, and including glorification of hateful entities in other categories depending on context due to their rare occurrence in our dataset.

Studies on religious hate

The problem of religious hate online and the design of approaches to detect and to counter it are relatively understudied. Among the few works that have focused only on religious targets, *Vidgen & Yasseri (2020)* present a novel dataset of 4,000 tweets annotated with three classes (*i.e.,* non-Islamophobic, weak Islamophobic and strong Islamophobic) and compare several classification algorithms and feature sets, arguing for a fine-grained classification of hate speech that goes beyond binary classes. *Chung et al. (2019)* present a large dataset of expert-curated sentence pairs containing Islamophobic messages and counter-narratives. This data has then been used to develop a platform that suggests responses to operators of non-governmental organizations fighting online hatred against Muslims (*Chung et al., 2021*). Our work is most similar to the study presented in *Albadi, Kurdi & Mishra (2018)*, introducing a dataset of 6,000 annotated tweets in Arabic, where six religious groups (*i.e.,* Jews, Christians, Sunnis, Muslims, Shia and Atheists) are represented with 1,000 tweets each. The work shows that half of the discussions about religion in the Arabic Twittersphere is hateful, targeting in most of the cases Jews, Atheists and Shia. The authors also compare different detection approaches, including lexicon-based, *n*-gram based, and deep learning algorithms. *Magdy et al. (2016)*, instead, deal with stances towards Muslims, collecting tweets about Islam after the Paris terrorist attacks in 2015 and analyzing whether they expressed a positive, negative or neutral stance. As regards antisemitism, *Zannettou et al. (2020)* quantify the escalation and spread of antisemitic memes and rhetoric across alt-right Web communities. They show that the use of antisemitic terms in Web communities is substantially influenced by real-world events like US Presidential elections. They do not present algorithms for antisemitic message detection but rather focus on lexicon-based analysis through word embeddings. Also, *Sabat, Canton-Ferrer & Giró-i-Nieto (2019)*

investigate hateful memes, considering Jew and Muslim memes, among others. Their detection experiments show that the visual cues are much more important than the linguistic ones when detecting hate speech memes.

Besides the works devoted exclusively to the analysis of religious hate online, a number of datasets include religion as one of the targets (for an overview, see *Poletto et al. (2021)* and *Vidgen & Derczynski (2020)*). The corpus described in *Vigna et al. (2017)* contains Facebook posts in Italian, where religion is one of the categories along with disability, politics, and gender, amongst others. Also, *Ishmam & Sharmin (2019)* present a Facebook corpus in Bengali where different annotated categories are provided, including religious hatred and religious comments. Along the same line, *Mossie & Wang (2020)* collect Facebook posts in Amharic covering different topics such as religion, politics, economy and ethnicity. *Olteanu et al. (2018)*, instead, focus on Reddit and Twitter data and analyze how attacks involving Arabs and Muslims trigger online hate speech, and in particular messages advocating violence. *Ousidhoum et al. (2019)* annotate tweets in English, French and Arabic along speech directness (*i.e.,* direct or indirect), hostility type and target, including religious affiliation.

Overall, existing works related to the classification of religious hate online present some limitations: first of all, Christianity is generally understudied, with the exception of *Albadi, Kurdi & Mishra (2018)*. We find, instead, that Christians are a rather common target, especially in English-written posts. Second, religion has been included in some of the existing taxonomies of hate targets, but no fine-grained categorization has been proposed in the past, distinguishing among different types of targets and hate speech forms, as proposed in our taxonomy. Third, religious hate online has been little explored from a cross-lingual perspective, although the country of origin has an impact on the perceived offensiveness of messages (*Salminen et al., 2018a*), and religious hate has country-specific cultural and historical roots. Indeed, our classification experiments and the following discussion aim at shedding light also on hate speech forms across religions and languages.

A TAXONOMY FOR RELIGIOUS HATE

Since abusive language online is a multi-faceted phenomenon and different categorizations of hate speech have been proposed over the years to account for different targets (*Kumar et al., 2018*; *Vidgen & Yasseri, 2020*; *Palmer et al., 2020*; *Zampieri et al., 2020*), we aim on the one hand to take the complexity of online abuse into account, and on the other hand to be compatible with existing annotation schemes as much as possible. We therefore build upon the hierarchical taxonomy proposed in *Zeinert, Inie & Derczynski (2021)*, which has been designed to annotate misogynist messages online but nevertheless provides a backbone for the fine-grained annotation of other target types. We chose *Zeinert, Inie & Derczynski (2021)*'s taxonomy as starting point since it captures a wide range of abuse types and targets, ultimately enabling a high degree of interoperability across annotated datasets. Indeed, although we focus on religious hate speech, we argue that providing annotations even for classes which are not of direct interest in this study (*e.g.*, misogyny, racism) allows data to be easily repurposed in future studies concerning other targets.[1]

[1] This is in contrast to the standard practice in datasets with focused targets, in which abusive posts that are not directed to the target under study are typically labeled as not abusive, thus undermining data interoperability for different use cases.

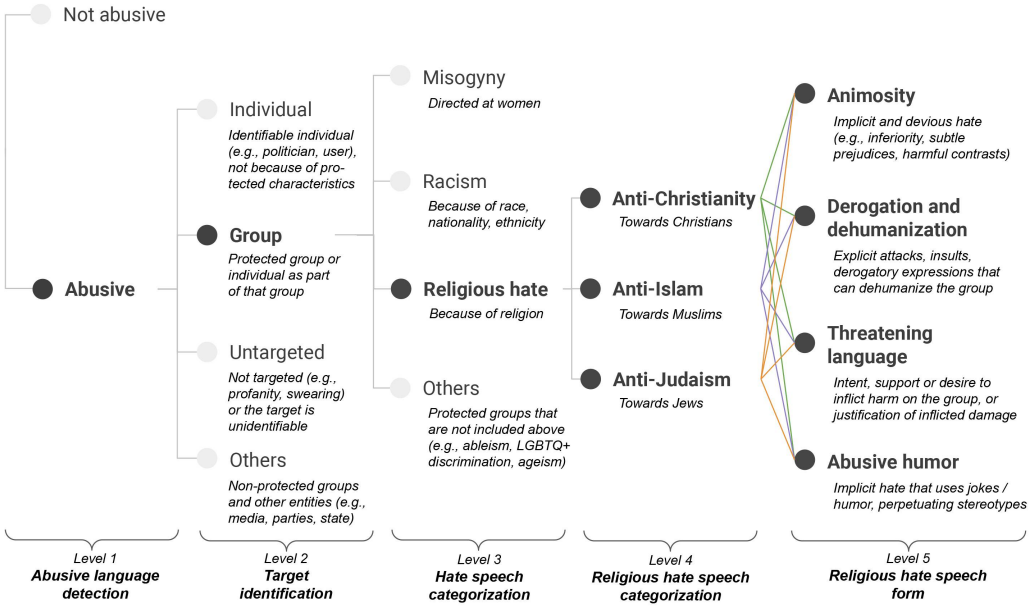


Figure 1 **Abusive language annotation taxonomy with a focus on religious hate.**
Full-size DOI: 10.7717/peerjcs.1128/fig-1

As shown in Fig. 1, we employ the first three levels of *Zeinert, Inie & Derczynski (2021)*'s taxonomy and extend the last one by adding a RELIGIOUS HATE category (cf. "Hate speech categorization" level in Fig. 1), which leads to two fine-grained levels for religious hate speech: (i) a level for fine-grained categorization of religious targets (cf. "Religious hate speech categorization" level in Fig. 1), and (ii) a level for religious hate speech forms (cf. "Religious hate speech form" level in Fig. 1). In the following, we first summarize the first three levels of the taxonomy. Then, we provide definitions and relevant examples for our proposed religious hate levels and associated labels.

Generic backbone of the taxonomy

In this section, we summarize the first three levels of the taxonomy introduced by *Zeinert, Inie & Derczynski (2021)*, which as mentioned above form the backbone of our annotation scheme for religious hate speech detection.

Level 1: Abusive language detection

The top-level of the taxonomy aims at distinguishing whether a post contains abusive language (ABUSIVE) or not (NOT ABUSIVE). We rely on the definition by *Caselli et al. (2020)* for abusive language classification, and consider as ABUSIVE any form of hurtful and derogatory language, discriminatory statement targeted at communities, groups or individuals based on personal characteristics, toxic comment, and untargeted profanity. This includes more subtle forms of abusive language, as detailed in the next sections.

Level 2: Target identification

Abusive posts can be further classified according to the target to which the abuse is directed. Consistently to *Zeinert, Inie & Derczynski (2021)*, possible labels for this level are GROUP, INDIVIDUAL, OTHERS and UNTARGETED:

- **GROUP**: the target of the abuse is a *protected* group or an individual as part of that group. We follow the widely-used definition of protected groups (*Röttger et al., 2021*; *Ramponi & Tonelli, 2022*; *Banko, MacKeen & Ray, 2020*), namely groups based on characteristics such as religion, ethnicity, race, gender identity, age, sex or sexual orientation, disability, and national origins. The category is related to *Davidson et al. (2017)*'s "hate speech" definition and focus on protected characteristics.
- **INDIVIDUAL**: the target of the abuse is an identifiable individual and can be either the user to whom one is replying to or a named person that is external to the conversation. The abuse directed to the individual is not because of a protected characteristic.
- **OTHERS**: the abuse is targeted at non-protected groups and entities, such as media, institutions, political parties, countries per se, and governments.
- **UNTARGETED**: the abuse is not targeted or the target is unidentifiable. Posts containing profanities, curses, swearing, and otherwise inappropriate language fall into this category.

Level 3: Hate speech categorization

Group-targeted abusive posts can be further classified based on the specific protected group the abuse is directed. We take MISOGYNY, RACISM and OTHERS from *Zeinert, Inie & Derczynski (2021)*, and further include a RELIGIOUS HATE category.

- **MISOGYNY**: the abuse expressed in the post is directed at women. This includes sexist content, harmful stereotypes, and subtle forms of abuse such as neosexism (*Zeinert, Inie & Derczynski, 2021*).
- **RACISM**: the characteristic to which the abuse is targeted is race, nationality or ethnicity of a group. This mainly includes explicit and implicit forms of racism and discrimination.
- **RELIGIOUS HATE**: the target of the abuse is a religious group or an individual because of its membership to that community. This includes abuse towards the three main monotheistic religions, namely Christianity, Islam and Judaism (see "Religious hate speech levels" for details).
- **OTHERS**: the abuse is directed at protected groups that are not included in the previous categories. These include ableism (discrimination against people with disabilities), LGBTQ+ discrimination (*i.e.*, on the ground of gender identity and sexual orientation), homophobia (discrimination against lesbian, gay, or bisexual people), transphobia (discrimination against transgender people) and ageism (age discrimination).

Religious hate speech levels

In this section, we present the proposed religious hate speech categories (level 4) and hate forms (level 5). We provide definitions along with relevant examples from our dataset in both English and Italian.

Level 4: Religious hate speech categorization

Abusive posts targeted to a religious group (RELIGIOUS HATE) can be further characterized based on the specific religion the abuse is directed. As detailed below, we include the main monotheistic religions as religious targets, namely ANTI-CHRISTIANITY, ANTI-ISLAM and ANTI-JUDAISM.

- **ANTI-CHRISTIANITY**: the target is Christianity and its faithful community. This category includes stereotypes aimed at Christians (*e.g.*, Christians portrayed as murderers or rapists), criticism of subgroups by other Christians, and general contempt for all or part of the Christian community.

  English example. "*christians and christianity actively enable sexual abuse environment... the "sexual scandal" involves actual human beings... but all you care about is his legacy and stupid teachings... fuck you very much [URL]*"

  Italian example. "*la storia del mio popolo in breve , le parole sono giuste e veritiere, quindi cari cristiani prima di fare le vittime studiate la vostra storia perché avete ancora le mani sporche di sangue [URL]*"—**English translation.** "*the history of my people in short , the words are right and truthful, so dear Christians before playing the victim study your history because your hands are still stained with blood [URL]*"

- **ANTI-ISLAM**: the abuse is targeted at Islam and its adherents. Common examples include implicit and veiled hatred, dehumanization of Muslims through explicit expressions, using the belonging to Islam as a reason to insult or criticize people's actions (*e.g.*, those of politicians), the willingness to inflict damage on Muslims or justifying and support already inflicted damages, stereotypes and prejudices (*e.g.*, Muslims as cutthroat), statements about the inferiority of the Islam belief compared to other religions, potentially abusive jokes, and Muslims as invaders.

  English example. "*world wake up before you accept muslims in your countries, those people they're not refugees, they're all terrorist who want to come to your state and make a plan how to attack, make sure to deport them all back.*"

  Italian example. "*l'#islam teme che la scuola gli remi contro. in realtà, l'islam sgozza e decapita i #cristiani in #ue. naturalmente i paladini pro multiculturalismo tutti zitti. [URL]*"—**English translation.** "*#islam fears that the school rows against it. actually, Islam slaughters and beheads #christians in #eu. naturally the pro multicultural champions all shut up. [URL]*"

- **ANTI-JUDAISM**: the abuse is directed at the Jewish community and its religion. Main examples include the denial of the holocaust and genocide, the explication of religious belief as justification for Jews' misbehavior, inappropriate comparisons between the Covid-19 pandemic (or vaccines) and the Shoah, insults through explicit swear words or comparisons with inhuman entities (mainly insects), abusive jokes and black humor, stereotypes and prejudices, the desire to inflict damage on the Jews or justification of a violence that has already been inflicted to them by others, and superiority claims at the expense of the Judaism belief.

English example. "*lmaoooo that's no what happened at all. jews believe that non-jews all a sort of animal created to be their slaves. they've treated everyone accordingly for thousands of years which is why so many people have kicked them out [URL]*"

Italian example. "*sicuramente babbo natale è un vecchio paziente e disponibile ma se in regalo gli chiedi di far sparire dalla terra certi 'parassiti' (ebrei, of course) il poveretto s'incavola e pure tanto! #antisemitismo*"—**English translation.** "*certainly santa claus is a patient and helpful old man but if you ask him to make certain 'parasites' (Jews, of course) disappear from the earth as a gift, the poor man gets pissed off a lot! #antisemitism*"

Level 5: Religious hate speech form

Abusive posts that are classified as targeted to a specific religious group on level 4 of the taxonomy are given an additional label indicating the way in which the hatred is manifested. We identify those labels starting from identity-directed abuse labels from *Vidgen et al. (2021)* (*i.e., derogation*, *animosity*, *threatening language*, *glorification of hateful entities* and *dehumanization*) and conducting a pilot annotation study on Italian data (see "Dataset creation"), which comprises several iterations and group discussions for refining the initial label set. Specifically, first iterations led to religion-specific labels that were iteratively handled and generalized to produce religion-agnostic labels, making the taxonomy more flexible and easier to apply across religious groups. A pivotal example is represented by the *diminishing* label for ANTI-JUDAISM, initially designed to describe instances of holocaust denial and inappropriate comparisons between the Covid-19 pandemic and the Shoah. We resolved this label by incorporating inappropriate contrasts as cases of ANIMOSITY (*i.e.,* implicit and soft hatred), whereas we treated holocaust denial differently according to the degree of hatred expressed in the post. We removed *glorification of hateful entities* from our initial label set including those rare cases in the other categories depending on context, and conceived an ABUSIVE HUMOR label to encompass instances of offensive jokes and black humor that could cause harm to target religious groups. Our final label set is thus largely based on *Vidgen et al. (2021)*'s labels and include ANIMOSITY and THREATENING LANGUAGE, merges *derogation* and *dehumanization* labels into a single DEROGATION AND DEHUMANIZATION category aiming at reducing ambiguity in annotation, further adding the rather understudied ABUSIVE HUMOR category. In the following we define and provide examples for each label of our final set.

- ANIMOSITY: the hatred expressed in the post is implicit and devious, nevertheless it is offensive and humiliating to the target. Implicit hate speech typically employs indirect or coded language (*ElSherief et al., 2021*), uses rhetorical devices to hide harmful meaning (*Caselli et al., 2020*), or subtly expresses negativity against certain groups (*Vidgen & Yasseri, 2020*). This makes identification of ANIMOSITY highly challenging for automated systems yet a fundamental part of abusive language detection (*ElSherief et al., 2021*; *Caselli et al., 2020*). In the context of religious hate speech, this category includes language in which: (a) the religious belief of the abuse target is made explicit, as if to mean that a wrong behavior is due to the religion to which the believer belongs; (b) the superiority of certain religions is manifested at the expense of others (*e.g.*, Christian

supremacy); (c) the religious group and its adherents are accused to receive "special and unfair treatments" compared to other groups; (d) a religion is used to criticize the actions of individuals (*e.g.*, politicians, the Pope) in geopolitical areas where another religion prevails; (e) ill-founded criticisms are argued against subgroups of the same religion or adherents to other religions; (f) stereotypes and prejudices about a religious group are perpetuated in a subtle manner without using explicit hate expressions; (g) identity words are used as insult through expressions that introduce presuppositions (*e.g.*, "They're Muslims/Jews/Christians after all"); (h) inappropriate harmful contrasts between the Covid-19 pandemic and the Shoah are made; (i) holocaust denial is manifested without explicitly insulting or demonizing the target group.

English example. "*[USER] we had the bible and christ before muhammed and the quran , aren't the latter rather superfluous ?*"

Italian example. "*agli ebrei tatuavano un numero oggi va il qr code. un passo avanti epocale. sono riusciti a farlo fare a loro stessi.*"—**English translation.** "*they tattooed Jews with a number today goes the qr code. an epochal step forward. they managed to get it done to themselves.*"

- DEROGATION AND DEHUMANIZATION: the hatred expressed against a target religious group is manifested in an explicit way. The target group is described in negative terms that can also dehumanize it (*e.g.*, adherents presented as animals or otherwise nonhuman entities), together with explicit attacks, insults or offensive and derogatory expressions (*e.g.*, "Islamic cutthroats"). Stereotypes and prejudices that lead to the demonization of the group also fall into this category. This represents one of the most direct and explicit form of hatred.

English example. "*[USER] [USER] #truth add: when muslims kill non-muslims - they stay silent too. hypocrites! wherever islam has infiltrated, it has brought about more harm than good. and we all know it. if it truly is a religion of peace, why has it done the most damage in the world? bunch of sickos! [URL]*"

Italian example. "*[USER] e islamica quindi: popolo sottosviluppato kattivo invidioso stupratore spacciatore con l'esclusiva del terrorismo.*"—**English translation.** "*she's Islamic so: underdeveloped bad people envious rapist drug dealer with the exclusive right of terrorism.*"

- THREATENING LANGUAGE: this category includes explicit forms of hateful language in which the willingness to inflict harm on a religious group is manifested, an already inflicted damage is justified or supported, or someone is encouraged to inflict such damage. In line with *Vidgen et al. (2021)*, harm includes physical and emotional violence, harassment and social exclusion. THREATENING LANGUAGE represents the most explicit and dangerous case of online hatred.

English example. "*[USER] [USER] from just muslim to ahmadi muslim, something new, something innovative, something never seen or heard before, nothing like the ahmadis had happened before, that means you made a new religion, let alone a sect, thats why you are apostates who deserve d3ath*"

Italian example. "*[USER] per combatterli bisogna mandarli a casa loro tutti i musulmani dico tutti buoni e cattivi non ne ne importa nulla cos risolviamo sicuramente il@problema e poi si decapitassero tra di loro*"—**English translation.** "*[USER] to fight them we must send them back to their home all the Muslims all good and bad it doesn't matter so we definitely solve the@problem and let them behead each other*"

- ABUSIVE HUMOR: an understudied case of implicit hate which uses jokes and black humor, indirectly leading to the stereotyping of target religious groups, and potentially perpetuating prejudices and causing harm to such groups. We treat it as a new and separate category since offensive jokes against religious groups are common in online fora, and the characteristics of such language are different from other cases of implicit language, being often sarcastic and metaphorical.

English example. "*i think my life is like a 12 yo boy in a catholic church cus man it do be sucking...*"

Italian example. "*[USER] la differenza tra un ebreo e una torta? il tempo di cottura*"—**English translation.** "*[USER] the difference between a Jew and a cake? cooking time*"

DATASET CREATION

In this section, we present the protocol we followed for collecting and annotating religious hate speech data in English and Italian. We then provide documentation in the form of data and artifacts statements (*Bender & Friedman, 2018*; *Ramponi & Tonelli, 2022*), as well as summary statistics and insights about the annotated corpus. While data collection follows the same protocol for both languages, we adopt two different approaches to data annotation.

For Italian, we had access to three native speakers with a background in computational linguistics, therefore annotation was performed following a protocol for experts that foresaw in-person discussion rounds and adjudication sessions.[2] Furthermore, the assignment of multiple labels was allowed to account for intersectionality when more than one target type was found in a tweet. For English, we could not follow the same approach, because we did not have the possibility to recruit English native speakers. However, we believe that being a native speaker should be a mandatory requirement for annotators to fully capture the subtleties and the cultural references in our domain of interest. We therefore resort to Amazon Mechanical Turk, following a standard workflow to collect multiple judgments from crowd-workers, while taking advantage of the EasyTurk tool (*Bocchi, Frasnelli & Aprosio, 2021*), that enables a strict quality control and monitoring of annotators' performance. In this case, however, the annotation protocol was slightly simplified, removing the possibility to assign multiple labels.

Although following exactly the same annotation process for the two languages would have been ideal to allow cross-lingual comparisons, mixing annotation approaches has already been investigated in prior work and has proven to have a limited impact on annotation quality (*Waseem, 2016*; *Sanguinetti et al., 2018*). In particular, *Waseem (2016)* found that cases of high-agreement provided by amateur annotators lead to relatively

[2] This was also motivated by the scarcity of Italian workers on crowdsourcing platforms such as Amazon Mechanical Turk.

good annotations as compared to expert annotators, making the crowd-sourcing protocol a viable approach in the case of complex multi-stage annotation schemes or absence of native speakers.

Data collection

We collected tweets for both English and Italian using the Twitter APIs by querying for key terms which are likely to occur in religious posts and refer to the three main monotheistic religions, namely Christianity, Islam and Judaism (cf. Appendix). The list of search keywords has been designed with and validated by domain experts involved in religious studies, as there were neither accessible nor complete lexicons to rely on at the time of collection. In order to make English and Italian portions of the dataset comparable, native or proficient English and Italian speakers curated term translations between the two languages.

We specifically avoid to use offensive words as part of our search terms since they are known to bias the dataset towards explicit rather than implicit abusive language (*Wiegand, Ruppenhofer & Kleinbauer, 2019*), and instead rely only on neutral terms with reference to specific religions, their branches, adherents, and sacred texts (*e.g.*, "Islam", "Catholicism", "Jewish", "Sunni", "Torah"). Although this has the downside of producing a lower proportion of abusive language in the resulting dataset (*Founta et al., 2018*), religion on online fora is a highly debated topic, and thus naturally leads to a high proportion of abusive language. Furthermore, to reduce as much as possible the diachronic bias of our dataset (*Florio et al., 2020*), we collected data spanning a nine-month-long time period (December 2020 – August 2021), thus mitigating the over-representation of online discourse about religion-related historical events.

Data collection has been performed in September 2021. After collection, we randomly sampled 3,000 Italian posts and 8,000 English posts for further abusive language annotation.

Data annotation

We devise annotation guidelines as described in "Taxonomy of religious hate" to ensure consistency in the labeling process across annotators and languages. Each post has been presented to annotators in an anonymized form and classified out-of-context with up to five labels (*i.e.,* until a leaf node is reached) corresponding to the levels of the taxonomy depicted in Fig. 1.

Italian subset. As mentioned above, the annotation for the Italian portion of the dataset has been performed by three expert annotators. All annotators are native speakers of Italian and have computational linguistics and computer science backgrounds. Two annotators identify themselves as females and one as male, with age ranges 20–30, 30–40 and 40–50. All annotators have a Catholic background, albeit being Atheists or non-practicing Christians. At the first stage of annotation, all annotators have been involved in discussion sessions to revise and validate the proposed labeling scheme and associated annotation guidelines. In order to enable intersectionality studies in the future, expert annotators provided multiple labels per instance during labeling, if applicable. To assess the quality of annotations, 20% of the dataset has been annotated by two annotators. We computed inter-annotator

agreement using Cohen's kappa (*Cohen, 1960*) for the two categorizations of interest to this work, namely *abusiveness* (*i.e.,* ABUSIVE *vs* NOT ABUSIVE), and *religion-related abusiveness* (*i.e.,* all posts with a religious hate label in level 3 of the taxonomy *vs* all others). We obtain $\kappa = 0.6530$ for abusiveness, and $\kappa = 0.5651$ for religion-related abusiveness before adjudication, which is moderate agreement. This is in line with other abusive language datasets (*Zeinert, Inie & Derczynski, 2021*) and reflects the difficulty of annotation for the task, which in our case was mainly due to a different interpretation of the tone of a given tweet. Then, cases of disagreement between the two annotators have been discussed and adjudicated by the third annotator. The resulting Italian portion of the dataset consists of 3,000 labeled examples.

English subset. The English portion of the dataset has been annotated by native English speakers from the United States, United Kingdom and Australia, following the annotation guidelines previously refined and validated for the Italian subset, using Amazon Mechanical Turk. Each tweet has been annotated by up to five annotators. To ensure high-quality annotations from non-expert crowd-workers and thus preventing potential scams, we follow *Leonardelli et al. (2021)*, and ask three expert linguists to annotate a batch of examples, then using tweets with perfect agreement as gold standard. We include a gold tweet in each group of ten examples to be annotated, using EasyTurk (*Bocchi, Frasnelli & Aprosio, 2021*). If a crowd-worker fails to evaluate the gold standard, we discard the remaining worker's annotations from our dataset. About 3% of the workers' annotations were removed from the study. Annotators were compensated on average with 8 US$ per hour. In order to make the annotation for the full taxonomy as clear and manageable as possible for the non-expert crowd, we proceed as follows. First, we ask crowd-workers to annotate the first two levels of the taxonomy (level 1 and 2 in Fig. 1). This allows us to identify potentially abusive messages towards religion (*i.e.,* those having majority GROUP label in level 2) to be further annotated. Second, annotators provide labeling decisions on GROUP-labeled tweets for the third and fourth level of the taxonomy (level 3 and 4 in Fig. 1). Although our quality control largely ensures spam annotations are not retained in the dataset, we argue that the subjective nature of the annotation task could lead to disagreement on a small fraction of previous decisions, and thus we give annotators the possibility to express their disagreement on previous annotations. We then remove from our final dataset the instances marked by the majority of annotators as disagreement. Third, we ask annotators to label the instances exhibiting ANTI-CHRISTIANITY, ANTI-ISLAM or ANTI-JUDAISM majority labels for the last level of the taxonomy. We include in the final dataset all examples having at least three individual annotations and a label majority on all non-leaf annotation stages. Furthermore, to minimize noisy annotations from the non-expert crowd, annotators have been instructed to only mark the main category at each annotation stage. As a result, the final English portion of the dataset consists of 7,028 fully labeled tweets.

We computed the inter-annotator agreement and use Krippendorff's alpha (*Krippendorff, 2004*) to account for multiple raters and sparse annotations. This led to $\alpha = 0.2667$ for *abusiveness*, which is in line with agreement reported on previous

datasets employing complex annotation schemes (*Ousidhoum et al., 2019*; *Sanguinetti et al., 2018*), and $\alpha = 0.8386$ for *religion-related abusiveness*, which is high agreement due to the multi-stage annotation workflow we devise for the English portion of the dataset. Since we adopt the strict quality-control protocol implemented through the EasyTurk application (*Bocchi, Frasnelli & Aprosio, 2021*), it is unlikely that disagreement is due to spammers and low-quality annotations, but it is rather due to different annotators' attitudes and background (*Aroyo & Welty, 2015*). Recent works advocate for reporting annotators' information when releasing datasets with toxicity annotation to account for biases and disagreement cases (*Sap et al., 2022*), but the processing of personal data concerning religious beliefs is explicitly prohibited by the EU General Data Protection Regulation (GDPR) (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e40-1-1), therefore we did not collect any such information from crowd-workers. For the sake of transparency, we release disaggregated labels for the full dataset, so to enable future research on annotators' disagreement.

Dataset documentation

In this section, we present relevant documentation for our dataset based on data and artifacts statements (*Bender & Friedman, 2018*; *Ramponi & Tonelli, 2022*).

Data statements

Data statements are provided in order to have a better overview of the rationale of creating the dataset and the taxonomy. According to *Bender & Friedman (2018)*, data statements represent a professional practice capable of "*provid[ing] context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software*". There are not only scientific implications involved, but ethical issues as well. Indeed, data statements for NLP alleviate issues related to exclusion and bias in language technology. In the following, we present data statements for the dataset we introduce in this work.

Curation rationale. The dataset was collected using neutral keywords related to religion (see Appendix) specifically avoiding offensive search terms that are known to bias the dataset towards explicit hate messages (*Wiegand, Ruppenhofer & Kleinbauer, 2019*). The dataset comprises an Italian and an English subset, and was created to the purpose of studying and mitigating the spread of online religious hate. A data instance is a tweet annotated according to a religious hate taxonomy scheme (see "Taxonomy for religious hate").

Language varieties. The variety of Italian (`it-IT`) and English (`en-*`, *i.e.,* without distinction between regional variants) represented in the dataset is spontaneous written speech.[3]

Speaker demographic. Data consists of anonymized posts, and thus user demographics are unknown.

[3] Many labels have been proposed to account for the characteristics of web varieties (*Guardiano, 2007*). All labels (*talky writing*, *conversational writing*, *electronic discourse*, etc.) agree with considering web varieties a hybrid form of communication with characteristics of both written and spoken language, to which such exclusive features as redundant punctuation, abbreviations, etc. are added.

Annotator demographic. *Italian subset.* Three native speakers of Italian, two females and a male, with age ranges 20–30, 30–40, and 40–50. All annotators have computational linguistics and computer science education, and a Catholic background (albeit being Atheists or non-practicing Christians). *English subset.* The pool of annotators are English speakers from the United States, United Kingdom and Australia. Other demographic characteristics have not been disclosed and are thus unknown.

Speech situation and text characteristics. The interaction is mainly asynchronous, the speaker's intended audience is everyone, and the text genre is social media with a focus on religious topics. Posts have been produced between December 2020 and August 2021 and collected in September 2021.

Preprocessing and data formatting. Posts have been anonymized by replacing user mentions and URLs with `[USER]` and `[URL]` placeholders, respectively. Texts have been preprocessed by lowercasing the text, removing newline characters, and unescaping possible HTML tags.

Artifacts statement

Lexical artifacts statement has been introduced by *Ramponi & Tonelli (2022)* as a way to document potential lexical biases when a dataset is released, providing a complementary view to data statements (*Bender & Friedman, 2018*). Lexical artifacts are defined as emergent correlations between tokens and labels in input data. We outline the statement for our dataset as follows.

Top lexical artifacts. In Table 1, we present the top $k = 15$ lexical artifacts for the abusive and religious hate classes for Italian and English data subsets.

Class definitions. The definition for abusive and religious hate classes are presented in "A taxonomy for religious hate" section.

Methods and resources. The $k = 15$ most informative tokens for the classes of interest have been extracted from the top features of a bag-of-words logistic regression model, in line with *Kennedy et al. (2020)*. Stopwords have been removed using the Stopwords ISO resource (https://github.com/stopwords-iso). Prior to computation, input texts have been preprocessed as described in "Preprocessing and data formatting". Tokens have been produced using the spaCy v3.3 tokenizer (https://spacy.io/) employing `it_core_news_sm` and `en_core_web_sm` models for Italian and English, respectively.

Dataset statistics

The final dataset consists of 10,028 annotated tweets, of which 3,000 are in Italian and 7,028 in English. The average tokens per post are 33.5 for Italian and 34.9 for English. Dataset statistics across taxonomy levels and languages are reported in Table 2. For Italian, we report between parenthesis the total count of annotations, given that it was possible to assign multiple labels to the same tweet. For English, numbers within parenthesis in level 5

Table 1 Top 15 most informative tokens for the ABUSIVE class (left) and RELIGIOUS HATE class (right) on Italian and English data subsets. Note that in some cases the lists capture some of the prejudices and stereotypes related to religion, see for example 'invasione' (invasion, mainly with respect to Muslims), 'finanza' (finance, mainly with respect to Jews), rape, terrorists, and '9/11'. Emojis credit: OpenMoji—the open-source emoji and icon project (CC BY-SA 4.0).

Rank	Abusive		Religious hate	
	Italian subset	English subset	Italian subset	English subset
1	Cazzo	Fuck	Islamica	Kill
2	Bergoglio	Fucking	Mica	Fuck
3	(*"Clown"*)	Kill	Islam	Cult
4	Fascisti	Fake	Invasione	Raped
5	Islamica	Ass	Portare	Jews
6	Invasione	Shit	Mussulmani	Jew
7	Mica	Rape	Maomettani	Representation
8	Merda	Stupid	Cazzo	Rape
9	Schifo	Hating	Nuovi	Terrorists
10	Sionisti	Representation	Islamici	Fake
11	(*"Poo"*)	Disgusting	Rabbino	Strange
12	Papa	Terrorist	Finanza	9/11
13	Islamici	Assholes	Clandestini	Cum
14	Dittatura	Biggest	(*"Poo"*)	Hypocrites
15	Vaccinati	Pedophiles	Islamico	Ugly

indicate the number of annotations considering multiple majority labels. The other count, instead, includes only tweets annotated with exactly one label.

Since the tweets were collected using the same set of keywords in the two languages, it is possible to some extent to compare the content of the two subsets. Overall, we observe that the proportion of abusive messages in Italian (31.8%) is much higher than in English (14.3%). Also the distribution of offenses against the three religions of interest is different: while in English they are more balanced (33% target Christianity, 40% Islam and 27% Judaism), in Italian most of the offenses are Islamophobic (16% are against Christians, 60% Muslims and 24% Jews). This would confirm the findings in *Ljujic et al. (2015)*, reporting that Italy is the most Islamophobic country in Europe after Hungary, and also the analysis provided by the Italian Observatory on Human Rights (http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-6/), showing that in 2021 the two groups most targeted on Twitter were women and Muslims. Concerning level 5, in both languages the cases of animosity are the most frequent, *i.e.,* mild or not overt abuse. Threats instead are the least frequent, also because they are prohibited by law in many countries, including Italy and UK.

In the Italian subset, annotators have often exploited the possibility to assign multiple labels to a tweet. We identify two main reasons for multiple assignments: (i) the presence of more than one target in the same message, for instance a politician and a religious group (see example 1 below), and (ii) a group targeted because of multiple sensitive attributes,

[4]Examples have been slightly edited.

Table 2 Dataset statistics across taxonomy levels and languages. Numbers outside parentheses indicate tweets having exactly that label, whereas numbers between parentheses are total counts per label.

	Class	Total	Italian	English
Level 1	ABUSIVE	1,961	954	1,007
	NOT ABUSIVE	8,067	2,046	6,021
Level 2	GROUP	1,130 (1,312)	404 (586)	726
	INDIVIDUAL	315 (412)	154 (251)	161
	UNTARGETED	115 (129)	48 (62)	67
	OTHERS	193 (323)	140 (270)	53
	multiple labels		208	–
Level 3	RELIGIOUS HATE	1,163 (1,220)	500 (557)	663
	MISOGYNY	10 (24)	7 (21)	3
	RACISM	42 (86)	14 (58)	28
	OTHERS	40 (51)	8 (19)	32
	Multiple labels		57	–
Level 4	ANTI-CHRISTIANITY	304 (313)	85 (94)	219
	ANTI-ISLAM	591 (603)	326 (338)	265
	ANTI-JUDAISM	310 (322)	131 (143)	179
	Multiple labels		15	–
Level 5	ANIMOSITY	720 (781)	369 (381)	351 (400)
	DEROGATION AND DEHUMANIZATION	139 (390)	139 (141)	196 (249)
	THREATENING LANGUAGE	35 (46)	17 (18)	18 (28)
	ABUSIVE HUMOR	51 (67)	17 (17)	34 (50)
	Multiple labels		15	–

for example Muslim women (see example 2 below).[4] The first cases are by far the most frequent.

(1) "*La Boldrini fedelissima dei talebani! Gli islamici la votano perché agevola il lavoro di islamizzazione alla velocità della luce!*"—**English translation.** "*Boldrini, the most loyal to the talibans! Islamics vote for her because she facilitates the work of Islamization at the speed of light!*"

(2) "*Le donne islamiche dicono che scegliendo tale religione allora scelgono anche di mettersi il velo e di avere relazioni solo con musulmani ma allora scegliete anche di andarvene aff**"—**English translation.** "*Islamic women say that by choosing this religion they also choose to wear the veil and have relations only with Muslims, but then you also choose to go f*ck yourself*"

EXPERIMENTS

In this section, we present the experimental setup and the classification methods we used for conducting experiments on the annotated dataset in both languages. We perform two binary classification tasks: one is aimed at abusive language detection (level 1 in Table 2), namely identifying posts that contain abusive language, and the second at religious

hate speech detection (level 3 in Table 2), *i.e.,* a fine-grained classification task aiming at detecting posts expressing religious hate.

Experimental setup

We cast both detection tasks as binary classification problems, in which the two classes to be predicted are ABUSIVE and NOT ABUSIVE for abusive language detection, whereas are RELIGIOUS HATE and NOT RELIGIOUS HATE for religious hate speech detection. For the sake of experiments, in religious hate speech detection we consider all tweets that do not exhibit a RELIGIOUS HATE label for the NOT RELIGIOUS HATE class. This leads to a total of 557 (18.57%) and 663 (8.26%) RELIGIOUS HATE examples for the Italian and English subsets, respectively. For abusive language detection, the number of ABUSIVE posts are 954 (31.80%) for Italian and 1007 (14.33%) for English. We preprocess all tweets by anonymizing user mentions and URLs with `[USER]` and `[URL]` placeholders, respectively. We then lowercase the text, remove newline characters, and unescape HTML tags.

Given the unbalanced distribution of labels across tasks and languages, we use macro-averaged precision (Prec), recall (Rec) and F1 score (F_1) as main metrics to assess the performance of our classifiers. This allows us to reliably evaluate performance of classification methods by giving the minority class equal importance to the majority one, and thus mitigating performance overestimation of other commonly used metrics (*e.g.*, accuracy (Acc), that we also present for the sake of completeness). For all our experiments, we use stratified k-fold cross-validation ($k = 5$) and report mean and standard deviation of all scores.

We perform experiments under two classification setups: a monolingual setup, in which all classifiers are trained and tested on in-language data, and a cross-lingual setup, in which classifiers are either (a) trained on both languages and tested on a target language (*i.e.*, cross-lingual learning), or (b) trained and tested on different languages (*i.e.*, zero-shot cross-lingual transfer). For the challenging yet more interesting cross-lingual setup, we employ multilingual pretrained language models, as described in the next section.

Classification algorithms

In order to gauge the level of difficulty of our tasks, we perform experiments using several classification methods, from naïve baselines and traditional machine learning classifiers to language-specific and multilingual pretrained language models.

Naïve baselines. We test three simple baselines to assess the complexity of both binary tasks: (i) always abusive (or always religious hate), which always predicts a tweet as ABUSIVE (or religious hate), (ii) always not abusive (or always not religious hate), which always predicts a tweet as NOT ABUSIVE (or NOT RELIGIOUS HATE), and (iii) random, which simply predicts a label at random. Although being trivial, those baselines serve as comparison for more complex classifiers, and thus enable to determine if more elaborate solutions are able to capture useful features to tackle both tasks.

Machine learning classifiers. We employ four traditional machine learning classifiers as implemented in the scikit-learn library (https://scikit-learn.org/stable/). Those algorithms

[5]OSCAR is a large-scale multilingual corpus based on filtered Common Crawl data that has been classified by language.

Table 3 Hyperparameter values used in our experiments for fine-tuning pretrained language models.

Hyperparameter	Value
Optimizer	AdamW
β_1, β_2	0.9, 0.99
Dropout	0.3
Epochs	3
Batch size	32
Learning rate (LR)	5e−5
LR scheduler	Slanted triangular
Decay factor	0.38
Cut fraction	0.2

are (i) decision tree, (ii) multinomial naïve Bayes, (iii) linear support vector classifier, and (iv) logistic regression. We use the scikit-learn count vectorizer for feature extraction and employ default hyperparameters as detailed in the official documentation across all our experiments. These baselines serve for comparison purposes to more recent solutions based on fine-tuning of monolingual or multilingual pretrained language models.

Language-specific pretrained language models. We experiment with transformer-based (*Vaswani et al., 2017*) language models pretrained on monolingual raw data, and fine-tune them on both binary tasks. For English, we employ (i) BERT (*Devlin et al., 2019*), whose pretraining corpus consists of 16GB of English text data (roughly 3.3B words) from BooksCorpus (*Zhu et al., 2015*) and English Wikipedia, and (ii) RoBERTa (*Liu et al., 2019*), whose pretraining procedure included 160GB of English raw text (*i.e.,* with additional news and web texts to BERT's pretraining corpus). We use the `bert-base-uncased` and `roberta-base` model versions, respectively, due to the relatively small size of our annotated datasets. We refer the reader to the original publications for details on different pretraining schemes of the two approaches (*Devlin et al., 2019*; *Liu et al., 2019*). For Italian, we use (i) AlBERTo (*Polignano et al., 2019*), a BERT-based language model pretrained on 191GB of text from a collection of tweets in Italian, and (ii) UmBERTo (`umberto-commoncrawl-cased-v1` version) (https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1), a RoBERTa-based language model pretrained on the deduplicated Italian portion of the OSCAR corpus (*Ortiz Suárez, Sagot & Romary, 2019*),[5] which accounts for a total of 70GB of raw text (rougly 11B words). We use all models as implemented in the MaChAmp v0.2 toolkit (*van der Goot et al., 2021b*) using default hyperparameters (Table 3), and a cross-entropy loss with balanced class weights to give equal importance to both classes.

Multilingual pretrained language models. For the sake of cross-lingual experiments, we employ two multilingual pretrained language models that include both Italian and English in the pretraining corpus: (i) multilingual BERT (*Devlin et al., 2019*), a BERT model pretrained on the 104 languages with the largest Wikipedia, and (ii) XLM-RoBERTa (*Conneau et al., 2020*), a RoBERTa-based model pretrained on 2.5TB of CommonCrawl

raw text containing 100 languages. We use the `bert-base-multilingual-cased` and `xlm-roberta-base` versions, respectively, as implemented in the MaChAmp v0.2 toolkit (*van der Goot et al., 2021b*) with default hyperparameters (Table 3). As for monolingual pretrained language models, we use cross-entropy loss with balanced class weights.

RESULTS AND ANALYSIS

In this section, we present monolingual and cross-lingual results and analysis of different classifiers across binary tasks and languages.

Monolingual classification setup

We here present the results under the monolingual setup, namely when training and testing classifiers on the same language (either the Italian or English subset).

Italian. In Tables 4 and 5 we report the k-fold cross validation results for abusive language detection and religious hate speech detection, respectively, on the Italian subset. Traditional machine learning classifiers generally improve the performance over naïve baselines, with logistic regression and linear support vector classifier being the most effective approaches of the category (67.21 and 66.23 F_1 score for abusive language detection, and 59.76 and 60.89 F_1 score for religious hate speech detection, respectively). Noticeably, multinomial naïve Bayes provides high precision scores, at the cost of a fairly low recall across tasks. However, best results for both tasks are obtained by fine-tuned language models, and specifically by monolingual models, with results ranging from 75.02 to 76.31 F_1 score for abusive language detection and from 64.86 to 65.69 F_1 score for religious hate speech detection. Interestingly, multilingual language models still outperform traditional machine learning classifiers according to the F_1 score metric. Indeed, multilingual BERT achieves 73.46 and 61.49 F_1 score, whereas XLM-RoBERTa reaches 71.37 and 63.90 F_1 score. Overall, the best results for both abusive language detection and religious hate speech detection are obtained by the fine-tuned UmBERTo model, which outperforms all other language models across all metrics. Although AlBERTo uses pretraining data pertaining to a domain that is closer to our dataset than UmBERTo (*i.e.,* Twitter texts), the consistent improvement of UmBERTo over AlBERTo suggests that the RoBERTa's pretraining scheme employed by UmBERTo is more beneficial on our tasks than the domain and size of the raw data used for pretraining.

English. In Tables 6 and 7 we report the k-fold cross validation results for abusive language detection and religious hate speech detection on the English portion of the dataset. Similarly to monolingual experiments on the Italian subset, language-specific pretrained language models provide the highest overall scores on both tasks. The best results according to macro-averaged precision, recall and F_1 score are achieved by RoBERTa, which consistently outperforms BERT achieving 72.56 F_1 score on abusive language detection and 64.94 F_1 score on religious hate speech detection. Results on the English subset, whose distribution is even more skewed than the Italian one towards the negative classes (*i.e.,* NOT ABUSIVE and NOT RELIGIOUS HATE), also reveal the weakness of the accuracy metric for assessing the

Table 4 ***k*-fold cross validation results for abusive language detection on the Italian subset.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest are in bold.

Method	Acc	Prec	Rec	F_1
Always abusive baseline	31.80 ± 0.1	15.90 ± 0.1	50.00 ± 0.0	24.13 ± 0.1
Always not abusive baseline	68.20 ± 0.1	34.10 ± 0.1	50.00 ± 0.0	40.54 ± 0.1
Random baseline	47.80 ± 2.9	48.11 ± 2.1	47.83 ± 2.5	46.17 ± 2.6
Decision tree	67.17 ± 1.1	61.10 ± 1.4	59.98 ± 1.1	60.31 ± 1.2
Multinomial naïve Bayes	74.47 ± 0.5	72.28 ± 0.9	64.19 ± 1.0	65.17 ± 1.2
Linear support vector classifier	72.10 ± 1.6	67.36 ± 2.1	65.64 ± 2.1	66.23 ± 2.2
Logistic regression	73.97 ± 1.4	70.00 ± 1.9	66.23 ± 1.7	67.21 ± 1.8
AlBERTo	77.20 ± 2.2	74.35 ± 2.1	76.57 ± 2.1	75.02 ± 2.2
UmBERTo	78.53 ± 1.1	**75.68** ± 1.0	**77.72** ± 1.4	**76.31** ± 1.1
Multilingual BERT	76.50 ± 2.4	73.17 ± 2.5	74.08 ± 3.0	73.46 ± 2.7
XLM-RoBERTa	72.83 ± 1.4	72.02 ± 1.3	74.86 ± 2.0	71.37 ± 0.9

Table 5 ***k*-fold cross validation results for religious hate speech detection on the Italian subset.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest are in bold.

Method	Acc	Prec	Rec	F_1
Always religious hate baseline	18.57 ± 0.1	9.28 ± 0.1	50.00 ± 0.0	15.66 ± 0.1
Always not religious hate baseline	81.43 ± 0.1	40.72 ± 0.1	50.00 ± 0.0	44.88 ± 0.1
Random baseline	48.83 ± 3.1	49.38 ± 1.9	48.98 ± 3.2	43.56 ± 2.7
Decision tree	77.60 ± 1.0	60.68 ± 2.4	58.88 ± 2.4	59.50 ± 2.4
Multinomial naïve Bayes	81.73 ± 0.4	**73.62** ± 7.9	51.57 ± 0.7	48.32 ± 1.2
Linear support vector classifier	79.53 ± 1.1	63.72 ± 2.1	59.78 ± 1.3	60.89 ± 1.5
Logistic regression	81.73 ± 0.6	68.17 ± 2.5	58.36 ± 1.3	59.76 ± 1.6
AlBERTo	72.77 ± 2.8	64.34 ± 1.1	71.08 ± 1.9	64.86 ± 1.8
UmBERTo	73.57 ± 2.1	64.87 ± 1.8	**71.86** ± 2.6	**65.69** ± 2.2
Multilingual BERT	68.83 ± 9.5	61.98 ± 5.5	67.77 ± 7.9	61.49 ± 8.2
XLM-RoBERTa	73.87 ± 1.9	63.08 ± 1.4	67.75 ± 3.1	63.90 ± 1.7

performance on our tasks. For instance, an "always not abusive" (or "always not religious hate") baseline seems to provide higher performance than most of the methods, albeit being unable to predict ABUSIVE (or RELIGIOUS HATE) examples. Overall, we notice lower performance on the English subset compared to the Italian subset despite the former being larger. We hypothesize this could be due to a more varied language use across users that write in English on Twitter, the different annotation methodology, or the more challenging distribution of classes in the dataset.

Cross-lingual classification setup

We here present the results under the cross-lingual setup, namely cross-lingual learning (*i.e.,* training multilingual models on both languages, then testing them on a target language), and zero-shot cross-lingual transfer (*i.e.,* training and testing multilingual models on different languages).

Table 6 ***k*-fold cross validation results for abusive language detection on the English subset.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest are in bold.

Method	Acc	Prec	Rec	F_1
Always abusive baseline	14.33 ± 0.1	7.16 ± 0.1	50.00 ± 0.0	12.53 ± 0.1
Always not abusive baseline	85.67 ± 0.1	42.84 ± 0.1	50.00 ± 0.0	46.14 ± 0.1
Random baseline	49.64 ± 1.3	50.21 ± 1.2	50.43 ± 2.4	42.67 ± 1.4
Decision tree	81.94 ± 0.4	59.61 ± 1.0	57.13 ± 0.9	57.93 ± 1.0
Multinomial naïve Bayes	85.91 ± 0.2	77.49 ± 6.2	51.30 ± 0.8	48.85 ± 1.5
Linear support vector classifier	83.32 ± 1.0	64.04 ± 2.4	61.11 ± 1.8	62.23 ± 2.1
Logistic regression	85.29 ± 0.6	67.42 ± 2.8	58.17 ± 1.3	59.98 ± 1.0
BERT	83.25 ± 1.5	69.55 ± 1.5	77.45 ± 1.9	72.01 ± 1.6
RoBERTa	83.25 ± 1.7	**69.96** ± 1.6	**78.77** ± 1.2	**72.56** ± 1.6
Multilingual BERT	78.40 ± 3.2	65.91 ± 1.6	75.57 ± 2.0	67.55 ± 2.3
XLM-RoBERTa	80.08 ± 1.4	67.32 ± 1.1	77.66 ± 1.0	69.58 ± 1.4

Table 7 ***k*-fold cross validation results for religious hate speech detection on the English subset.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest are in bold.

Method	Acc	Prec	Rec	F_1
Always religious hate baseline	9.43 ± 0.1	4.72 ± 0.1	50.00 ± 0.0	8.62 ± 0.1
Always not religious hate baseline	90.57 ± 0.1	45.28 ± 0.1	50.00 ± 0.0	47.52 ± 0.1
Random baseline	49.16 ± 1.6	49.81 ± 1.0	49.43 ± 2.9	39.61 ± 1.4
Decision tree	86.64 ± 0.5	54.29 ± 1.0	52.90 ± 0.9	53.21 ± 1.0
Multinomial naïve Bayes	90.52 ± 0.1	55.29 ± 20.	50.04 ± 0.2	47.66 ± 0.3
Linear support vector classifier	87.07 ± 0.2	57.55 ± 1.3	55.56 ± 1.5	56.23 ± 1.5
Logistic regression	89.68 ± 0.5	61.03 ± 4.6	53.02 ± 1.7	53.46 ± 2.5
BERT	81.87 ± 1.7	62.70 ± 0.8	74.52 ± 1.3	64.93 ± 1.0
RoBERTa	81.46 ± 1.4	**62.71** ± 1.0	**75.38** ± 1.9	**64.94** ± 1.3
Multilingual BERT	75.95 ± 2.5	59.68 ± 0.6	72.40 ± 0.7	60.05 ± 1.5
XLM-RoBERTa	78.53 ± 4.0	60.20 ± 1.2	71.19 ± 3.9	61.15 ± 2.1

Cross-lingual learning. We investigate the ability of multilingual pretrained language models (namely, multilingual BERT and XLM-RoBERTa) on abusive language detection and religious hate speech detection for a target language – either Italian (*IT*) or English (*EN*) – when trained on both language subsets (*i.e.,* *IT+EN*). This allows us to empirically determine if more data, even if it belongs to different languages, provides useful signals to models in the training process. Consistently to monolingual experiments, target language data is partitioned into $k = 5$ 80%/20% train/test splits, and each model is thus evaluated k times on the k test portions. As shown in Table 8 (top), for abusive language detection in English (*IT+EN* → *EN*) both fine-tuned language models improve the F_1 performance over the respective variants trained on English language only (cf. Table 6). The improvement is as large as 2.79 F_1 score for multilingual BERT and 1.52 F_1 score for XLM-RoBERTa. When testing those models on the Italian subset (*IT+EN* → *IT*), however, we notice only a small

Table 8 ***k*-fold cross validation results for cross-lingual abusive language detection and religious hate speech detection.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest and setups are in bold.

Setup	Method	Acc	Prec	Rec	F_1
		Abusive language detection			
IT+EN → *EN*	Multilingual BERT	82.03 ± 1.2	68.01 ± 1.3	75.83 ± 1.5	70.34 ± 1.4
	XLM-RoBERTa	82.95 ± 1.8	**68.99** ± 1.3	**75.95** ± 2.4	**71.10** ± 1.1
IT+EN → *IT*	Multilingual BERT	72.90 ± 3.2	72.00 ± 2.2	75.15 ± 2.4	71.63 ± 2.9
	XLM-RoBERTa	72.93 ± 1.5	**72.38** ± 1.7	**75.74** ± 2.0	**71.81** ± 1.6
		Religious hate speech detection			
IT+EN → *EN*	Multilingual BERT	80.49 ± 1.7	60.97 ± 0.9	**71.53** ± 1.3	**62.70** ± 1.3
	XLM-RoBERTa	82.60 ± 4.1	**61.83** ± 1.1	69.36 ± 7.1	62.55 ± 2.1
IT+EN → *IT*	Multilingual BERT	68.03 ± 1.6	**63.30** ± 0.6	**71.36** ± 1.0	**62.09** ± 1.1
	XLM-RoBERTa	66.60 ± 3.3	62.41 ± 1.8	69.93 ± 2.9	60.71 ± 2.7

improvement for XLM-RoBERTa (0.44 F_1 score), whereas a performance degradation occurs for multilingual BERT (−1.83 F_1 score) (cf. Table 4). Regarding religious hate speech detection (Table 8 (bottom)), we notice a similar trend in the performance of models. Specifically, results on English test data (*IT+EN* → *EN*) show an improvement of 2.65 F_1 score for multilingual BERT and 1.40 F_1 score for XLM-RoBERTa (cf. Table 7), whereas results on Italian test data (*IT+EN* → *IT*) confirm that multilingual models struggle in generalizing across languages on the task (cf. Table 5). Multilingual BERT slightly improves the performance when training on additional English data (0.60 F_1 score), whereas XLM-RoBERTa exhibits substantially lower scores (−3.19 F_1 score). Overall, monolingual models provide better performance across tasks and setups (cf. Tables 4, 5, 6 and 7), suggesting that multilingual models can be primarily leveraged in *zero-shot cross-lingual* setups, as described in the next section.

Zero-shot cross-lingual transfer. In real-world setups, there is typically a lack of annotated data for a particular task on a target language. Since our dataset contains two language subsets, we simulate a *zero-shot cross-lingual* setup, assuming we only have data from a given *source language* and thus aiming at classifying data on a *target language*. In Table 9 we report the results of our experiments across tasks (*i.e., abusive language detection* and *religious hate speech detection*) and languages (*i.e., IT* → *EN* and *EN* → *IT*). Overall, we notice that XLM-RoBERTa provides the best overall performance across our languages and tasks. By a closer look, zero-shot *abusive language detection* results on English data (*IT* → *EN*) reach 63.56 F_1 score, whereas zero-shot results on Italian data (*EN* → *IT*) achieve 70.57 F_1 score (Table 9 (top)). This is remarkable, especially for XLM-RoBERTa on the *EN* → *IT* setup, whose performance are near to the XLM-RoBERTa variant trained on in-language data (−0.80 F_1 score) and not too far from results of the in-language UmBERTo language model (−5.74 F_1 score) (cf. Table 4). As regards *religious hate speech detection*, XLM-RoBERTa results under the *EN* → *IT* zero-shot setup compared to results when using in-language data exhibits a 3.63 F_1 score degradation (cf. Table 5), whereas for the *IT* → *EN* setup the performance drop is even smaller (*i.e.,* −2.62 F_1 score) (cf. Table 7).

Table 9 ***k*-fold cross validation results for zero-shot cross-lingual abusive language detection and religious hate speech detection.** Results are averages (with standard deviation as subscript). Precision, recall, and F_1 scores are macro-averages. Best results for metrics of interest and setups are in bold.

Setup	Method	Acc	Prec	Rec	F_1
		Abusive language detection			
IT → *EN*	Multilingual BERT	85.05 ± 0.4	**66.49** ± 2.3	58.97 ± 3.9	60.33 ± 4.8
	XLM-RoBERTa	78.27 ± 5.2	64.66 ± 2.0	**69.24** ± 7.3	**63.56** ± 3.1
EN → *IT*	Multilingual BERT	73.23 ± 1.7	69.42 ± 2.5	66.96 ± 2.5	67.41 ± 2.1
	XLM-RoBERTa	75.40 ± 1.4	**71.75** ± 1.8	**69.96** ± 1.2	**70.57** ± 1.2
		Religious hate speech detection			
IT → *EN*	Multilingual BERT	82.71 ± 3.9	56.61 ± 2.7	58.09 ± 3.3	56.79 ± 3.0
	XLM-RoBERTa	84.35 ± 3.6	**58.87** ± 1.4	**60.54** ± 4.8	**58.53** ± 3.1
EN → *IT*	Multilingual BERT	73.63 ± 2.4	59.68 ± 2.0	61.50 ± 3.8	59.96 ± 2.5
	XLM-RoBERTa	72.20 ± 3.3	**60.20** ± 1.0	**63.11** ± 3.6	**60.27** ± 0.6

Noticeably, the difference between XLM-RoBERTa on a zero-shot setup compared to the best monolingual model is −6.41 F_1 score for the *IT* →*EN* setup, and −5.42 F_1 score for the *EN* → *IT* setup. This suggests that zero-shot cross-lingual learning is a viable option for both abusive language detection and religious hate speech detection when annotated data on a target language is unavailable. This finding is in contrast with previous works showing that zero-shot settings do not yield satisfying results for cross-lingual hate speech detection (*Nozza, 2021*). Our experiments suggest that adopting exactly the same sampling procedures in terms of query terms and time period makes the subsets of tweets in the two languages comparable, and this similarity is probably captured well by multilingual transformer models.

DISCUSSION

In this section we provide a quantitative and qualitative analysis of annotated tweets in the last two levels of the taxonomy (Fig. 1), specifically investigating religious hate forms across religions and languages. Finally, we present a discussion on limitations of our study and directions for future work.

Analysis of religious hate speech forms across religions and languages

In Table 10 we present the proportion of religious hate speech forms across languages and religions in our dataset. By analyzing data qualitatively, we observe some differences in the distribution of hate forms depending on the targeted religion. Islam appears to be the religion with the greatest number of DEROGATION AND DEHUMANIZATION tweets across languages (30.8% for Italian and 39.4% for English). It is generally portrayed as hostile, dangerous and threatening for the West, or mentioned within the narration of immigration and integration. Muslim hatred is especially evident in the Italian subset (Table 2). We hypothesize that the main reason for this different distribution is that the Italian portion of the dataset contains tweets that are mainly produced by a single nation, where the majority of citizens declares to be Catholic (https://www.uaar.it/doxa2019/), whereas the English subset comprises tweets produced by speakers from all over the world. For this

Table 10 Distribution of religious hate speech forms across languages and religions.

	Italian			English		
	Christianity	Islam	Judaism	Christianity	Islam	Judaism
ANIMOSITY	71.4%	63.1%	77.9%	63.2%	56.2%	56.4%
DEROGATION AND DEHUMANIZATION	25.3%	30.8%	12.1%	29.2%	39.4%	27.6%
THREATENING LANGUAGE	2.2%	3.8%	4.3%	1.9%	2.8%	4.3%
ABUSIVE HUMOR	1.1%	2.3%	5.7%	5.7%	1.6%	11.7%

reason, we argue that it is also possible to find more systematic hateful patterns in the Italian tweets. For instance, 'islamico' (Islamic) has started being used as a noun with a negative connotation. Moreover, only in the Italian data subset the reference to Islam is often exploited to criticize politicians and the Pope. Overall, the distribution of hate labels in our dataset reflects the anti-Muslim sentiment that is on the rise across Western countries, and whose growth has been analyzed by social scientists in recent work (*Cervi, Tejedor & Gracia, 2021*).

Hatred towards Jews, on the contrary, appears to be more implicit and subtle across languages. Specifically, Judaism is the religion that exhibits the largest proportion of ABUSIVE HUMOR tweets in both Italian and English. A highly common pattern in the Italian data is the occurrence of inappropriate juxtapositions between Covid-19 themes (*e.g.*, vaccines) and the Shoah, which we treat as ANIMOSITY in our annotation guidelines. Prejudices, either overtly malevolent (*i.e.*, DEROGATION AND DEHUMANIZATION) or ostensibly benign (*i.e.*, ANIMOSITY) (see *Wilson (1996)* for more details), are also relatively frequent in both languages.

Finally, DEROGATION AND DEHUMANIZATION tweets are more frequent in ANTI-CHRISTIANITY tweets than ANTI-JUDAISM ones, and slightly less frequent than ANTI-ISLAM posts. Moreover, according to a qualitative exploration of our labeled dataset, reference to Christianity is often used to affirm religion supremacy at the expense of other religions. Overall, we notice that the distribution of religious hate speech targeting Christianity is similar across languages, with Italian tweets expressing more ANIMOSITY and English tweets containing a larger fraction of ABUSIVE HUMOR.

Limitations and future work

We acknowledge some limitations of this study. First, our dataset focuses on the three main monotheistic religions of the world, adopting a Western-centric approach. In future work we aim to extend our dataset to additional religions following our religious hate speech annotation scheme, and further include other languages from diverse families. Second, our annotations are provided out-of-context, however a fraction of the posts could be more reliably labeled if the thread context is taken into account in the annotation process. We aim to investigate the impact of contextual information in future work, along the line of *Vidgen et al. (2021)* and *Menini, Aprosio & Tonelli (2021)*. Third, we acknowledge that our dataset could embed subtle biases due to both the data sampling procedure and annotators' background. For data collection, even if we designed search keywords with experts in religious studies, deliberately avoiding to rely on offensive word lists,

our sampling procedure is still purposive (*Wiegand, Ruppenhofer & Kleinbauer, 2019*). Since a more desirable random sampling is often impractical due to the low presence of abusive posts on Twitter (*Founta et al., 2018*), in future work we aim to investigate alternative, half-way data sampling techniques such as keyword expansion after seed terms bootstrapping to further diversify our dataset. As regards annotation, the Italian portion of the dataset has been labeled by expert annotators with Christianity background due to difficulties in finding expert annotators who are both native speakers of Italian and whose belief is Judaism or Islam. However, annotators' lived experiences, sensibility and faith play a role in how they could perceive religious hate. In order to ultimately enable to take into account as many annotators' perspectives as possible in religious hate speech detection, we decided to release disaggregated annotations in our dataset. We believe this could enable research directions in modeling different annotators' perspectives, following successful applications in subjective tasks (*Davani, Díaz & Prabhakaran, 2022*; *Ramponi & Leonardelli, 2022*), as well as smoothly provide valuable extensions to our dataset.

Interesting future avenues for research include also studying the impact of lexical normalization on downstream abusive language detection and religious hate speech detection performance, using monolingual (*van der Goot et al., 2020*; *Baldwin et al., 2015*) or multilingual datasets (*van der Goot et al., 2021a*), as well as exploiting multiple annotations on the Italian portion of the dataset to study intersectionality.

CONCLUSIONS

In this work we present an in-depth analysis of religious hate as expressed in Twitter data in Italian and English. We first introduce a taxonomy with five different annotation layers, which has been designed to be fully compatible with the taxonomy for misogyny annotation presented in *Zeinert, Inie & Derczynski (2021)*. The taxonomy covers the three main monotheistic religions as well as different types of offense, from the more subtle ones (ANIMOSITY) to the very explicit THREATENING LANGUAGE. Based on this categorization, we collect English and Italian tweets using neutral religion-related terms, and annotate them using two different approaches: for English, we rely on a standard workflow collecting crowd-worker judgments through Amazon Mechanical Turk, whereas for Italian the tweets are annotated by three native speakers going through discussion and adjudication sessions when needed. Experts are also given the possibility to assign multiple labels to the same tweets, opening future avenues for research on intersectionality. Using the resulting dataset, we perform experiments by comparing different classification algorithms in two binary tasks: the detection of abusive language, and the identification of religious hate. Our results show that monolingual pretrained language models provide the best performance in both tasks, and that zero-shot cross-lingual transfer is a viable option for religious hate speech detection when annotated data for the target language is not available.

APPENDIX

Keywords

English: Islam, muslim, muslims, moslem, moslems, islamic, islamics, quran, qur'an, koran, sunnah, hadith, athar, sunni, sunnism, shia, shi'a, shiism, shi'ism, sunnis, sunnite, sunnites, shias, shi'as, shiite, shi'ite, shiites, shi'ites, judaism, jew, jews, jewish, jewishes, jewry, jewries, tanach, hebrew bible, torah, midrash, talmud, rabbanite, rabbinic, rabbinism, rabbinicism, karaite, qaraite, karaism, qaraism, rabbanites, rabbi, rabbis, rabbinist, rabbinists, rabbinicist, rabbinicists, karaites, qaraites, christianity, christians, bible, old testament, new testament, catholicity, catholicism, catholic church, protestantism, eastern orthodox church, orthodox catholic church, catholic, catholics, protestant, protestants, orthodox, orthodoxes.

Italian: Islam, musulmano, musulmana, mussulmano, mussulmana, musulmani, musulmane, mussulmani, mussulmane, maomettano, maomettana, maomettani, maomettane, islamico, islamica, islamici, islamiche, corano, quran, qur'an, koran, sunna, sunnah, hadith, athar, sunnismo, sciismo, sunnita, sunniti, sciita, sciiti, ebraismo, giudaismo, ebreo, ebrea, ebrei, ebree, ebraico, ebraica, ebraici, ebraiche, tanakh, tenakh, bibbia ebraica, torah, torà, midrash, talmud, rabbinico, rabbinismo, caraita, karaita, karaismo, karaitismo, caraismo, caraitismo, rabbino, rabbina, rabbini, rabbine, rabbinista, rabbinisti, rabbiniste, caraiti, karaiti, cristianesimo, cristiani, bibbia, antico testamento, vecchio testamento, primo testamento, antica alleanza, nuovo testamento, nuova alleanza, cattolicità, cattolicesimo, chiesa cattolica, protestantesimo, chiesa ortodossa orientale, chiesa ortodossa d'oriente, chiesa cristiana orientale, chiesa cristiana d'oriente, chiesa cattolica ortodossa, cattolico, cattolica, cattolici, cattoliche, protestante, protestanti, ortodosso, ortodossa, ortodossi, ortodosse.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the PROTECTOR European project (ISFP-2020-AG-PROTECT-101034216-PROTECTOR). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
PROTECTOR European project: ISFP-2020-AG-PROTECT-101034216-PROTECTOR.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Alan Ramponi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Benedetta Testa conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Sara Tonelli conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Elisabetta Jezek conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available on Zenodo: Ramponi, Alan, Testa, Benedetta, Tonelli, Sara, & Jezek, Elisabetta. (2022). Resources for the article: "Addressing religious hate online: From taxonomy creation to automated detection" (v1.0). Zenodo. https://doi.org/10.5281/zenodo.7082068.

REFERENCES

Akiwowo S, Vidgen B, Prabhakaran V, Waseem Z (eds.) 2020. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics.

Albadi N, Kurdi M, Mishra S. 2018. Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Piscataway: IEEE, 69–76.

Anzovino ME, Fersini E, Rosso P. 2018. Automatic identification and classification of misogynistic language on Twitter. In: *International conference on applications of natural language to data bases*.

Aroyo L, Welty C. 2015. Truth is a lie: crowd truth and the seven myths of human annotation. *AI Magazine* **36(1)**:15–24 DOI 10.1609/aimag.v36i1.2564.

Awan I. 2017. Cyber-extremism: ISIS and the power of social media. *Society* **54**:1–12 DOI 10.1007/s12115-016-0108-3.

Baldwin T, De Marneffe MC, Han B, Kim Y-B, Ritter A, Xu W. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In: *Proceedings of the Workshop on Noisy User-Generated Text*. Beijing: Association for Computational Linguistics, 126–135.

Banko M, MacKeen B, Ray L. 2020. A unified taxonomy of harmful content. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, 125–137.

Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M. 2019. SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 54–63.

Bender EM, Friedman B. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* **6**:587–604 DOI 10.1162/tacl_a_00041.

Bocchi L, Frasnelli V, Palmero Aprosio A. 2021. EasyTurk: a user-friendly interface for high-quality linguistic annotation with Amazon Mechanical Turk. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 106–112.

Boyd D, Golder S, Lotan G. 2010. Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences, HICSS '10*. Washington: IEEE Computer Society, 1–10.

Breitfeller L, Ahn E, Jurgens D, Tsvetkov Y. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, 1664–1674.

Carmona MÁÁ, Guzmán-Falcón E, Montes-y Gómez M, Escalante HJ, Pineda LV, Reyes-Meza V, Sulayes AR. 2018. Overview of MEX-A3T at IberEval 2018: authorship and aggressiveness analysis in Mexican Spanish tweets. In: *Proceedings of IberEval 2018*. 74–96.

Caselli T, Basile V, Mitrović J, Kartoziya I, Granitzer M. 2020. I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 6193–6202.

Cervi L, Tejedor S, Gracia M. 2021. What kind of Islamophobia? Representation of Muslims and Islam in Italian and Spanish media. *Religions* **12(6)**:427 DOI 10.3390/rel12060427.

Charitidis P, Doropoulos S, Vologiannidis S, Papastergiou I, Karakeva S. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media* **17**:100071 DOI 10.1016/j.osnem.2020.100071.

Christopherson KM. 2007. The positive and negative implications of anonymity in internet social interactions: "On the internet, nobody knows you're a dog". *Computers in Human Behavior* **23(6)**:3038–3056 DOI 10.1016/j.chb.2006.09.001.

Chung Y-L, Kuzmenko E, Tekiroğlu SS, Guerini M. 2019. CONAN—Counter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 2819–2829.

Chung Y-L, Tekiroğlu SS, Tonelli S, Guerini M. 2021. Empowering NGOs in countering online hate messages. *Online Social Networks and Media* **24**:100150 DOI 10.1016/j.osnem.2021.100150.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20(1)**:37–46 DOI 10.1177/001316446002000104.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, 8440–8451.

Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology* **20(2)**:10:1–10:22.

Davani AM, Díaz M, Prabhakaran V. 2022. Dealing with disagreements: looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* **10**:92–110.

Davidson T, Warmsley D, Macy M, Weber I. 2017. Automated hate speech detection and the problem of offensive language. In: *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, Montreal, Canada*.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis: Association for Computational Linguistics, 4171–4186.

ElSherief M, Ziems C, Muchlinski D, Anupindi V, Seybolt J, De Choudhury M, Yang D. 2021. Latent hatred: a benchmark for understanding implicit hate speech. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 345–363.

Fersini E, Nozza D, Rosso P. 2020. AMI EVALITA2020: automatic misogyny identification. In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765 of CEUR Workshop Proceedings, Online*.

Fersini E, Rosso P, Anzovino M. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), volume 2150 of CEUR Workshop Proceedings*. Spain: Sevilla,.

Florio K, Basile V, Polignano M, Basile P, Patti V. 2020. Time of your hate: the challenge of time in hate speech detection on social media. *Applied Sciences* **10(12)**:4180.

Fortuna P, Nunes S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys* **51(4)**:1–85:30.

Founta A-M, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, Kourtellis N. 2018. Large scale crowdsourcing and characterization

of Twitter abusive behavior. In: *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

Garimella K, Weber I, De Choudhury M. 2016. Quote RTs on Twitter: usage of the new feature for political discourse. In: *Proceedings of the 8th ACM Conference on Web Science, WebSci '16*. New York: Association for Computing Machinery, 200–204.

Gray KL, Buyukozturk B, Hill ZG. 2017. Blurring the boundaries: using gamergate to examine real and symbolic violence against women in contemporary gaming culture. *Sociology Compass* **11(3)**:e12458 DOI 10.1111/soc4.12458.

Grimminger L, Klinger R. 2021. Hate towards the political opponent: a Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online. Association for Computational Linguistics, 171–180.

Guardiano C. 2007. Aspetti dell'italiano parlato. *Journal of Pragmatics* **39(2)**:431–435 DOI 10.1016/j.pragma.2006.08.001.

Ishmam A, Sharmin S. 2019. Hateful speech detection in public facebook pages for the Bengali language. In: *18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*. 555–560.

Johnson NF, Leahy R, Restrepo NJ, Velásquez N, Zheng M, Manrique P, Devkota P, Wuchty S. 2019. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* **573(7773)**:261–265 DOI 10.1038/s41586-019-1494-7.

Kennedy B, Jin X, Davani AMostafazadeh, Dehghani M, Ren X. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, 5435–5442.

Krippendorff K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research* **30(3)**:411–433.

Kumar R, Ojha AK, Lahiri B, Zampieri M, Malmasi S, Murdock V, Kadar D (eds.) 2020. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille: European Language Resources Association (ELRA).

Kumar R, Ojha AK, Malmasi S, Zampieri M. 2018. Benchmarking aggression identification in social media. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe: Association for Computational Linguistics, 1–11.

Larimore S, Kennedy I, Haskett B, Arseniev-Koehler A. 2021. Reconsidering annotator disagreement about racist language: noise or signal? In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online. Association for Computational Linguistics, 81–90.

Lee M, Hwalbin K, Okhyun K. 2015. Why do people retweet a tweet? Altruistic, egoistic, and reciprocity motivations for retweeting. *Psychologia* **58(4)**:189–201 DOI 10.2117/psysoc.2015.189.

Leonardelli E, Menini S, Palmero Aprosio A, Guerini M, Tonelli S. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 10528–10539.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. *CoRR* ArXiv preprint. arXiv:abs/1907.11692.

Ljujic V, Engel V, Castriota A, Carter A. 2015. Xenophobia, radicalism and hate crime in Europe 2015. Technical Report HDIM.NGO, Organization for security and co-operation in Europe.

Magdy W, Darwish K, Abokhodair N, Rahimi A, Baldwin T. 2016. ISISisNotIslam or DeportAllMuslims?: Predicting unspoken views. In: *Proceedings of the 8th ACM Conference on Web Science, WebSci '16*. ACM, 95–106.

Menini S, Aprosio AP, Tonelli S. 2021. Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection. *CoRR* ArXiv preprint. arXiv:abs/2103.14916.

Mossie Z, Wang J-H. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management* **57(3)**:102087 DOI 10.1016/j.ipm.2019.102087.

Nozza D. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online. Association for Computational Linguistics, 907–914.

Nozza D, Bianchi F, Hovy D. 2021. HONEST: measuring hurtful sentence completion in language models. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, 2398–2406.

Olteanu A, Castillo C, Boy J, Varshney K. 2018. The effect of extremist violence on hateful speech online. In: *Proceedings of the International AAAI Conference on Web and Social Media, volume 12*.

Onabola O, Ma Z, Yang X, Akera B, Abdulrahman I, Xue J, Liu D, Bengio Y. 2021. hBERT + BiasCorp fighting racism on the web. In: *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Kyiv: Association for Computational Linguistics, 26–33.

Ortiz Suárez PJ, Sagot B, Romary L. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In: *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 9–16.

Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung D-Y. 2019. Multilingual and multi-aspect hate speech analysis. In: *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 4675–4684.

Palmer A, Carr C, Robinson M, Sanders J. 2020. Cold: annotation scheme and evaluation data set for complex offensive language in English. *Journal for Language Technology and Computational Linguistics* **34**:1–28.

Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* **55(2)**:477–523 DOI 10.1007/s10579-020-09502-8.

Polignano M, Basile P, De Gemmis M, Semeraro G, Basile V. 2019. AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: *6th Italian Conference on Computational Linguistics, CLiC-it 2019*. CEUR, 1–6.

Ramponi A, Leonardelli E. 2022. DH-FBK at SemEval-2022 task 4: leveraging annotators' disagreement and multiple data views for patronizing language detection. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, 324–334.

Ramponi A, Tonelli S. 2022. Features or spurious artifacts? Data-centric baselines for fair and robust hate speech detection. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 3027–3040.

Ranasinghe T, Zampieri M. 2022. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing* **21(1)**:4:1–4:13.

Röttger P, Vidgen B, Nguyen D, Waseem Z, Margetts H, Pierrehumbert J. 2021. HateCheck: functional tests for hate speech detection models. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 41–58.

Sabat BO, Canton-Ferrer C, Giró-i-Nieto X. 2019. Hate speech in pixels: detection of offensive memes towards automatic moderation. *CoRR*. ArXiv preprint. arXiv:1910.02334.

Salminen JO, Almerekhi H, Milenkovic M, Jung S-G, An J, Kwak H, Jansen BJ. 2018b. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: *Twelfth International AAAI Conference on Web and Social Media*.

Salminen J, Veronesi F, Almerekhi H, Jung S-G, Jansen BJ. 2018a. Online hate interpretation varies by country, but more by individual: a statistical analysis using crowdsourced ratings. In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. Piscataway: IEEE, 88–94.

Sánchez-Junquera J, Chulvi B, Rosso P, Ponzetto SP. 2021. How do you speak about immigrants? Taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences* **11(8)**:3610 DOI 10.3390/app11083610.

Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M. 2018. An Italian Twitter corpus of hate speech against immigrants. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).

Sap M, Card D, Gabriel S, Choi Y, Smith NA. 2019. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 1668–1678.

Sap M, Swayamdipta S, Vianna L, Zhou X, Choi Y, Smith NA. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle: Association for Computational Linguistics, 5884–5906.

Thompson R. 2011. Radicalization and the use of social media. *Journal of Strategic Security* **4(4)**:167–190 DOI 10.5038/1944-0472.4.4.8.

van der Goot R, Ramponi A, Caselli T, Cafagna M, De Mattei L. 2020. Norm it! lexical normalization for Italian and its downstream effects for dependency parsing. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 6272–6278.

van der Goot R, Ramponi A, Zubiaga A, Plank B, Muller B, San Vicente Roncal I, Ljubešić N, Ö Çetinoğlu, Mahendra R, Çolakoğlu T, Baldwin T, Caselli T, Sidorenko W. 2021a. MultiLexNorm: a shared task on multilingual lexical normalization. In: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Online. Association for Computational Linguistics, 493–509.

van der Goot R, Üstün A, Ramponi A, Sharaf I, Plank B. 2021b. Massive choice, ample tasks (MaChAmp): a toolkit for multi-task learning in NLP. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online. Association for Computational Linguistics, 176–197.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems 30*.

Vidgen B, Derczynski L. 2020. Directions in abusive language training data, a systematic review: garbage in, garbage out. *PLOS ONE* **15(12)**:e0243300 DOI 10.1371/journal.pone.0243300.

Vidgen B, Nguyen D, Margetts H, Rossini P, Tromble R. 2021. Introducing CAD: the contextual abuse dataset. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, 2289–2303.

Vidgen B, Yasseri T. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics* **17(1)**:66–78 DOI 10.1080/19331681.2019.1702607.

Vigna FD, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. 2017. Hate me, hate me not: Hate speech detection on Facebook. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), volume 1816 of CEUR Workshop Proceedings*. Venice: 86–95. CEUR-WS.org.

Waseem Z. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin: Association for Computational Linguistics.

Waseem Z, Chung WHK, Hovy D, Tetreault JR (eds.) 2017a. In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver: Association for Computational Linguistics.

Waseem Z, Davidson T, Warmsley D, Weber I. 2017b. Understanding abuse: a typology of abusive language detection subtasks. In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver: Association for Computational Linguistics, 78–84.

Waseem Z, Hovy D. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL Student Research Workshop*. San Diego: Association for Computational Linguistics, 88–93.

Wiegand M, Ruppenhofer J, Kleinbauer T. 2019. Detection of abusive language: the problem of biased datasets. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: Association for Computational Linguistics, 602–608.

Wilson TC. 1996. Compliments will get you nowhere: Benign stereotypes, prejudice and anti-semitism. *The Sociological Quarterly* **37(3)**:465–479 DOI 10.1111/j.1533-8525.1996.tb00749.x.

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. 2019. SemEval-2019 task 6: identifying and categorizing offensive language in social media (OffensEval). In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis: Association for Computational Linguistics, 75–86.

Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Ç Çöltekin. 2020. SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020). In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 1425–1447.

Zannettou S, Finkelstein J, Bradlyn B, Blackburn J. 2020. A quantitative approach to understanding online antisemitism. In: *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*. American Association for Artificial Intelligence (AAAI) Press, 786–797.

Zeinert P, Inie N, Derczynski L. 2021. Annotating online misogyny. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online. Association for Computational Linguistics, 3181–3197.

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S. 2015. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway: IEEE, 19–27.

PeerJ Computer Science

Critical reflections on three popular computational linguistic approaches to examine Twitter discourses

Dan Heaton, Jeremie Clos, Elena Nichele and Joel Fischer

School of Computer Science, University of Nottingham, Nottingham, United Kingdom

ABSTRACT

Although computational linguistic methods—such as topic modelling, sentiment analysis and emotion detection—can provide social media researchers with insights into online public discourses, it is not inherent as to how these methods should be used, with a lack of transparent instructions on how to apply them in a critical way. There is a growing body of work focusing on the strengths and shortcomings of these methods. Through applying best practices for using these methods within the literature, we focus on setting expectations, presenting trajectories, examining with context and critically reflecting on the diachronic Twitter discourse of two case studies: the longitudinal discourse of the NHS Covid-19 digital contact-tracing app and the snapshot discourse of the Ofqual A Level grade calculation algorithm, both related to the UK. We identified difficulties in interpretation and potential application in all three of the approaches. Other shortcomings, such the detection of negation and sarcasm, were also found. We discuss the need for further transparency of these methods for diachronic social media researchers, including the potential for combining these approaches with qualitative ones—such as corpus linguistics and critical discourse analysis—in a more formal framework.

Subjects Computational Linguistics, Data Mining and Machine Learning, Network Science and Online Social Networks, Text Mining, Sentiment Analysis

Keywords Critical reflection I, Sentiment analysis, Topic modelling, Emotion detection, Covid-19, Tweet, Contact-tracing

Submitted 31 May 2022
Accepted 19 December 2022
Published 30 January 2023

Corresponding author
Dan Heaton,
daniel.heaton@nottingham.ac.uk

Academic editor
Arkaitz Zubiaga

Additional Information and Declarations can be found on page 31

DOI **10.7717/peerj-cs.1211**

Copyright
2023 Heaton et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Computational linguistics lend themselves to study public discourses around topics of significant current societal concern. With most cultural phenomena, social media websites such as Twitter host a plethora of views relating to current affairs, and there is interest in gathering these views to make judgements on the general trends of opinion expressed (*McCormick et al., 2017*). Many offer opinions on this public-access site, providing a large dataset that can be analysed by means of open APIs in real-time (*Kumar, Morstatter & Liu, 2014*). A popular way to explore the views expressed on social media is to use 'off-the-shelf' computational linguistic methods such as topic modelling, sentiment analysis and emotion detection. This is because these methods may be less intrusive and more cost-effective when compared with interviews or experiments (*Rout et al., 2018*).

Social media researchers have used methods such as topic modelling, sentiment analysis and emotion detection to mine social media and online spaces for views expressed about

various topics. Cases of interest include research by *Hu, Chancellor & De Choudhury (2019)* into discourses relating to homelessness on social media, where topic modelling techniques were deployed to investigate common thematic threads between those who identify as homeless and those who do not. Sentiment analysis and emotion detection techniques have been explored by *Tang et al. (2020)* and *Wang et al. (2021)*, who both used them to garner greater insights into discourses relating to online and remote education.

However, there is now a growing body of work involving topic modelling, sentiment analysis and emotion detection that focuses on critically examining how these methods can be applied and the relative success of the methods. There have been studies that focus on the existing limited critical examination of particularly the shortcomings of these methods when applied to large corpora drawn from social media. Some initial evaluative work has been done in the studies by *Jiang, Brubaker & Fiesler (2017)*, who critiqued the binary sentiment scale used in sentiment analysis packages as having little tangible meaning about the views expressed in text. Additionally, emotion modelling critiques by social media researchers include work by *Fast, Chen & Bernstein (2016)* and their comparison of numerous models to develop a better performing function.

Therefore, to further contribute to this discourse, we perform an exploratory critical reflection in this article. We do this by following an analysis process that draws on best practices. We then draw out instances from our analysis of Twitter data that demonstrate the application of these methods and how they have been used to capture meaning from the data. By using the critically reflective weather model by *Maclean (2016)* as the basis for this, we look at how this could aid social media researchers in the use of these methods.

Two examples of public discourses around topics of societal concern have been chosen to aid this critical reflection. The first, the Covid-19 contact-tracing application, was launched in the UK with the intention to easily identify users who have come into contact with someone who has tested positive for Covid-19, based on Bluetooth proximity technology. The app has had widespread impact on the lives of citizens of the country such as by informing them to self-isolate at home for up to 10 days (*Kretzschmar et al., 2020*). It is estimated that at the height of the recent wave the app has sent more than 600,000 self-isolation alerts in a single week in July 2021 alone (https://www.bbc.co.uk/news/technology-57929162). Related to our research, Twitter was one of the websites used to explore attitudes to digital contact-tracing in Germany, where issues of privacy were among the most intensely discussed topics (*Arzt, Poller & Vallejo, 2021*). The second example, the Ofqual algorithm used for calculating A Level grades in England and Wales in 2020, was initially brought in to replace examinations in the Covid-19 pandemic. However, public outcry saw the algorithm scrapped and teacher assessment grades used instead. These examples were chosen due to having contrasting timelines (an 11 month period for the COVID-19 app and a three week period for the Ofqual algorithm).

The aims of this article is to critically reflect on the use of topic modelling, sentiment analysis and emotion detection when analysing social media data. We do this *via* following a set of collated best practices from within the field. Our work does not aim to provide an exhaustive analysis of the two case studies on social media more broadly, but to examine two focused datasets from a singular source (Twitter) to offer critical reflections on the

application of computational linguistic methods. As a result of this, we offer a set of reflections to be used by social media researchers when undertaking further inquiries into online discourses that could benefit from using these methods.

RELATED WORK

Herein, we introduce the computational methods we chose to examine our Twitter datasets, review related work in the realm of social media analysis and consider existing approaches to addressing shortcomings in these approaches. This review will uncover existing best practices and motivate how we combine these to create our analytical process. We also provide a brief overview of the digital contact tracing in the UK and the Ofqual A Level algorithm for context.

Computational linguistic methods

Three popular computational linguistic tools have been chosen as the means the explore two public discourse case studies on Twitter. The three approaches—topic modelling, sentiment analysis and emotion detection—are united in their use of language to either describe or make predictions about a corpus. Thus, they are descriptive and predictive in their function. They are furthermore popular choices when there is a large amount of linguistic data to explore. Topic modelling is best used to uncover latent topics present within large bodies of text (*Nikolenko, Koltcov & Koltsova, 2017*). Sentiment analysis—otherwise known as opinion mining—uses predictive algorithms on a binary polarity scale to provide insight into the views expressed in text (*Liu, 2010*; *Vyas & Uma, 2018*). Emotion detection methods use similar predictive detection algorithms to sentiment analysis to ascertain emotions or states of being that may be expressed within a text (*Sailunaz et al., 2018*).

We are also aware of a number of other techniques that could have been used, but have chosen to focus on these three due to scope, suitability and accessibility. Ultimately, these three chosen methods have consistently been used by researchers within this realm of social media discourses. They have also been chosen as their scores can be compared in a diachronic way.

Topic modelling

Automated topic modelling, particularly Latent Dirichlet Allocation (LDA), is seen as beneficial in qualitative text studies due to its focus on uncovering underlying topics present within a series of documents (*Nikolenko, Koltcov & Koltsova, 2017*). The use of LDA as a topic modelling method with Twitter data has grown in interest in recent years (*Arianto & Anuraga, 2020*) and various techniques have been designed to undertake this investigation. There has been focus on using bigrams—pairs of adjacent words—to form topic models through LDA when investigating views expressed on Twitter (*B.Srinivasan & Kumar, 2019*). This is closely related to the work done by *Yang & Zhang (2018)*, who combined this with the Bag-of-Words (BoW) model. This aided the creation of reliable topic models when working with Twitter data, especially when concerned with context dependence of short texts. *Prihatini et al. (2018)* combined, compared and contrasted LDA with Term

Frequency Inverse Document Frequency (TF-IDF), a method regularly used for feature extraction of texts, when examining online news and related media articles. They found that the Precision, Recall and F-Measure values of LDA was higher than TF-IDF for predicting topics, thus recommending it as the more suitable method.

The gensim package is a popular choice for topic modelling and LDA, using the analysis of co-occurrence patterns to identify latent structures in plain text documents (*Rehùøek & Sojka, 2011*). *Hidayatullah et al. (2019)* used gensim as the LDA model to investigate topics and trends regarding climate change and weather on Twitter, where five key topics were able to be defined through this method. Another relevant study was by *Song et al. (2019)*, who investigated topics in media discourses regarding illegal compensation given to victims of occupational injuries. The topics discovered were then used as recommendation points for ensuring efficient occupational health and safety schemes protect vulnerable employees from illegal practices, exemplifying the practical applications that using LDA can have.

This method has been been used within contemporary social media studies. *Hu, Chancellor & De Choudhury (2019)* used gensim as their topic modelling tool when investigating discourses relating to homelessness on social media. They looked first at the the blog posts of those who identify as homeless on the blogging site tumblr and compared these to the blog posts of those who do not identify as homeless. By using LDA, thirteen latent topics were identified as part of the discourse of homeless blog posts and seventeen topics were uncovered to be part of the control group's discourse. Within this study, LDA was deemed successful in its identification of the different ways in which homeless people and non-homeless people discuss the topic of homelessness. It was recommended that organisations can tap into these topical discourses to raise awareness and promote support, exemplifying the power of using this computational linguistic method. Additionally, *Sengupta (2019)* also used LDA to investigate latent topics in subreddit forums. This combined manual inspection with LDA to validate findings.

It has been documented about how to best find the appropriate number of topics for a dataset when using LDA. In their exploration of NLP-based techniques, *Nguyen et al. (2020)* suggested that the needs of the researcher must be examined: a small number of topics for a broad overview, and more topics for finer detail.

There have been critiques of topic modelling also. *Maier et al. (2018)* questioned the validity and reliability of LDA and offered an evaluative framework that enabled communication researchers to more effectively deploy this method. Their four recommendations were categorised as pre-processing the data, selecting parameters carefully, evaluating the reliability of the model and, finally, checking the validation of the results *via* manual review. We aim to extend this idea to other computational linguistic methods *via* our own examination and reflection. Researchers have also explored the limitations of 'off-the-shelf' topic modelling. In particular, stemmers (which are used to conflate several words to a shared meaning in topic modelling) were critiqued by *Schofield & Mimno (2016)*, suggesting that they had no effect on the outcomes of topic models and often disrupted topic stability. Contemporary Twitter studies such as *Saura, Ribeiro-Soriano & Saldaña (2022)* also expressed that labelling LDA topics is a manual process and, as such, this introduced bias into their results.

Sentiment analysis

Another popular investigation method to uncover the views expressed on social media is sentiment analysis. Sentiment analysis is defined as 'the computational treatment of opinions, sentiments and subjectivity of text' (*Medhat, Hassan & Korashy, 2014*). In its common form, it uses a binary polarity scale from negative to positive, with neutral in between, initially examining lexicon individually with a view to providing an overall sentiment of a text (*Liu, 2010*). It is often used to easily ascertain data that will provide insight into the opinions of others (*Vyas & Uma, 2018*), which enables investigation through a predictive element. Sentiment analysis has been used to investigate meaning of Twitter discourses for this same reason (*Liu & Zhang, 2012*). For this particular study, we chose to focus on dictionary-based approaches as these were the most common approaches seen in similar studies. Nevertheless, there are other approaches, such as CNN or deep learning methods, that are becoming popular within NLP.

A staple of most dictionary-based sentiment analysis methods is the Naive Bayes classifier, which makes the simplifying assumption that all words are sampled independently (hence the 'naive'). This produces a conditional probability of a document belonging to a category (*Rish et al., 2001*). *Pak & Paroubek (2010)* built a classifier based on Naive Bayes principles that classified Twitter data about a range of topics according to traditional sentiment polarity, but claimed it outperformed other comparable classifiers in accuracy. *Ulfa et al. (2018)* combined the Naive Bayes classifier with a Mutual Information method when examining tweets relating to tourism in Lombok and yielded a high classification accuracy.

Two popular sentiment analysis modules that have been used for Twitter opinion mining are TextBlob and VADER. TextBlob also uses the Naives Bayes model and also provides a subjectivity measure (*Loria, 2018*). *Pokharel (2020)* used this technique to analyse the response to the COVID-19 outbreak in Nepal and found that majority of the people of Nepal were taking a positive and hopeful approach, but there were instances of fear, sadness and disgust exhibited too. Additionally, *Sivalakshmi et al. (2021)* explored the sentiment towards the COVID-19 vaccine using TextBlob and concluded that the discourse was neutral-to-negative in polarity. An important factor here was that they identified that TextBlob was unable to read tokenized special characters as a limitation of the module, and factored this into their analysis. *Tang et al. (2020)* adopted the TextBlob sentiment software within their ConceptGuide system. The sentiment analysis here played a crucial element in the evaluation of their tool and future work proposes a learning efficiency analysis that may use similar principles, exemplifying the utilisation of TextBlob outside of a social media context.

VADER classification module acts in a similar way to TextBlob, using bigrams to attempt to detect negation in syntactical structures. Various studies have ranged from mining emotions from online video comments (*Chaithra, 2019*) to looking specifically at sentiment on Twitter. Chauhan et al. recognised the changing and challenging formats of tweets might have had an impact on previous work undertaken, and, because of VADER's sensitivity to social media formats, it is a more suitable module to use and yielded more accurate results (*Chauhan, Bansal & Goel, 2018*). *Mustaqim et al. (2020)* combined VADER's with

the k-nearest neighbour algorithm, and found that the two of them together yielded greater insight into the Twitter discourse concerning Indonesian forest fires in 2019. Because of its suitability for social media research, VADER was used by *Park, Ciampaglia & Ferrara (2016)* to investigate fashion trends on Instagram. Additionally, VADER has applications beyond social media discourse research. It was used in the study about student perceptions of a virtual teaching assistant in a study by *Wang et al. (2021)*.

Gupta & Joshi (2021) looked specifically at the role of negation in Twitter sentiment analysis, with a focus is on negation scope detection and negation handling methods. Their main conclusion is that negation is not a trivial task but entails many challenges, such as implicit negations and negation exception cases. By exploring this within the healthcare domain on Twitter, they have been able to lay foundations for this to be applied to other areas of research on Twitter, stressing the importance of handling negation exception cases, where negation cue act as non-cue, through a deep learning model. In similar fashion, sentiment analysis modules have been critiqued for the lack of consideration for sarcasm. *González-Ibánez, Muresan & Wacholder (2011)* investigated this by comparing machine learning techniques to human reviewers, and found that neither performed well in classifying sarcastic tweets. *Villena-Román & Garcıa-Morera (2013)* built on this when suggesting that sentiment analysis is so complex that humans will often disagree on the sentiment of a text in question.

There have been developments to the standard sentiment analysis models in recent years, with social media researchers going beyond the standard approach. For example, *Watanabe (2021)* set forth an alternative to sentiment analysis entitled Latent Semantic Scaling (LSS), which used principles of latent semantic analysis (word embeddings) to improve the traditional sentiment model. In addition, *van Atteveldt, van der Velden & Boukes (2021)* deployed a survey of many different sentiment analysis techniques and concluded that sentiment dictionaries were not of acceptable standards of validity and, while machine and deep learning outperformed dictionary-based methods the best performance was attained by trained human coding. This further shows that work to combine quantitative and qualitative methods here may be of use. All the studies exemplify that, to make best use of these computational language analysis methods, they must be combined with other approaches to validate results.

Some studies have begun to offer ideas about how to overcome the limitations of sentiment analysis. Some of these have been technical, such as *Ribeiro et al. (2020)* evaluating the number of actionable bugs through using an agnostic testing methodology for NLP models. However, some have focused on interpretation. For example. the study by *Agarwal et al. (2015)* offered the idea of using contextual information alongside sentiment results to better interpret them. This goes alongside other suggestions, such as the research into presenting sentiment over a period of time as a trajectory by *Howard (2021)*. This allows researchers to see how sentiment trends begin to form and develop, thus offering sentiment change as an interpretation strand. These ideas were important to us when considering our approach to analysis.

Emotion detection

Emotion detection from text could be seen as a sister method to sentiment analysis in the sense that it attempts to assign to documents a multidimensional vector representing its emotional valence (resp. sentiment valence) across a set of pre-defined emotion categories (resp. sentiment polarity), based on observation of text. One of the earliest pieces of work from the profile of mood states set forth by *Bollen, Mao & Pepe (2011)*. They used a psychometric instrument to extract six mood states (tension, depression, anger, vigor, fatigue, confusion) from the aggregated Twitter content and found that social, political, cultural and economic events are correlated with significant, even if delayed fluctuations of mood.

A popular contemporary emotion detection module for Python is EmoLex. The algorithm takes a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), which was manually done initially by crowdsourcing (*Mohammad & Turney, 2013*). This model has been applied to Twitter investigations, such as the analysis by *Aribowo & Khomsah (2021)* of Indonesian Twitter users' response to the COVID-19 pandemic. *Mathur, Kubde & Vaidya (2020)* again used a similar process to find high levels of trust and fear in tweets relating to COVID-19 from all over the world. Facebook data has also been investigated using EmoLex, with notable examples being the two 2019 studies by Balakrishnan et al. that investigated the emotions in the online diabetes community. EmoLex was refined using string-based Multinomial Naïve Bayes algorithm, with results indicating a 6.3% improvement (*i.e.*, 82% *vs.* 75.7% for average F-score) when compared to the EmoLex alone (*Balakrishnan et al., 2019*; *Balakrishnan & Kaur, 2019*).

The examples here illustrate the benefits of using EmoLex and emotion detection models within a social media context. However, within most of these studies in a social media context, there has been limited focus on the shortcomings of using these methods. Some studies have commented on the need for more complex emotional categories (*Jiang, Brubaker & Fiesler, 2017*), which may solve overlap between categories (*De Silva et al., 2018*; *Leung et al., 2021*). Also, some have attempted to mitigate human review biases seen when classifying texts as 'neutral' because of annotator's uncertainty around the best-fit category, rather than it actually being neutral (*Fujioka et al., 2019*). Additionally, EmoLex was used by *Fast, Chen & Bernstein (2016)* to develop Empath, a tool that can generate and validate new lexical categories on demand from a small set of seed terms to use of large texts. This offered some critique of EmoLex, as they stated that it correlates imperfectly with Linguistic Inquiry and Word Count analytical procedures. This is an important starting point for a more open, critically reflective discussion on the practical advantages and limitations of using each of these computational linguistic methods when investigating social media discourses. A reflection on the use of these methods may benefit those in the future that are looking to deploy them as part of their investigation.

Summary and research gap

Overall, this review shows that these computational methods are used regularly in the social media research community, although exploring them critically through the medium

of reflection is a newer discourse. While there are studies that have used these methods, or similar ones or variations, alongside qualitative methods, few have been formalised so that they include existing best practices and critically reflective ideas. Therefore, this contribution will combine existing best practice ideas through an analytical process that encourages critical reflections.

Context of public discourse case studies

We chose two recent case studies of public discourses as examples for our work. These two—the NHS COVID-19 digital contact tracing app and the Ofqual A Level algorithm—were chosen due to their significant societal concern in the United Kingdom and their contrasting timelines (a dataset spanning eleven months for the COVID-19 app and a dataset spanning three weeks for the Ofqual algorithm). This section introduces some background context of these two case studies.

Digital contact tracing during COVID-19 in the UK

NHS COVID-19, the contact-tracing application created by Serco on behalf of the UK government to track active cases of COVID-19, has impacted daily life in the United Kingdom since its launch (*Kretzschmar et al., 2020*). The application is available on mobile phones and uses exposure logging, developed by Apple and Google (*NHS England, 2021*). This technology allows the application to send alerts, using a randomly generated identification number, when the user is near to another application user who has logged a positive COVID-19 test. Some of the issues users encountered were backwards incompatibility, incorrect alerts and false positive tests, which meant that users had to self-isolate for ten days when the result was incorrect, impacting income and well-being (*Kent, 2020*). *Mbwogge (2021)* claims that a symptom-based contact-tracing system failed to meet the testing and tracing needs in the United Kingdom, which is further evidenced by the fact that cases of and deaths relating to COVID-19 increased to be the highest in Europe. Although adoption of the application was encouraged, the uptake of the application was less than expected at 20.9 million downloads, with no specific figures published on active users.

There are a growing number of research projects investigating the public attitudes towards digital contact-tracing in the UK. *Williams et al. (2021)* interviewed 27 participants over online video conferencing before the release of the COVID-19 app in the UK and found the response to be mixed and heavily influenced by moral reasoning. Analysis revealed five themes: lack of information and misconceptions surrounding COVID-19 contact-tracing apps; concerns over privacy; concerns over stigma; concerns over uptake; and contact-tracing as the 'greater good'. *Samuel et al. (2021)* conducted 35 semi-structured qualitative interviews in April 2020, showing interviewees' views about the potential of app for contact tracing. Participants showcased a range of misconceptions and worries. However, as there was no follow-up to this study, it was impossible to discover which of the participants would then choose to download or not to download the app once it was launched in September 2020.

After the app launched, *Dowthwaite et al. (2021)* found through surveying 1,001 UK adults that half of participants installed the app and a high number of these participants

complied with the app on a regular basis. They also found that there were issues surrounding trust and understanding that hindered the effective adoption of the app.

In July 2021, media focus was placed on the app once more due to the amount of alerts to self-isolate being sent to users. An increased amount of positive COVID-19 cases in the UK and the relaxation of government restrictions meant that users of the app were more frequently contacted due to exposure logging (*Abbasi, 2021*). This was negative perceived in the press and, as a result, the pejorative portmanteau 'pingdemic' was coined (*Rimmer, 2021*).

The Ofqual A level algorithm in England and Wales

Ofqual's use of an algorithm for calculating Advanced Level qualification grades in the summer of 2020 was a highly contested issue (*Kelly, 2021*). The algorithm was deployed to replace the standard A Level qualification, which had been abandoned due to the COVID-19 pandemic. Instead, the algorithm used data such as prior centre attainment and postcodes and factored these into the generated grade for each qualification. This led to a number of students receiving grades that were different to their predicted outcomes submitted by their teachers when the results were released to students on August 13th 2020.

Despite defending the use of the algorithm initially, due to public outcry, the UK Government retracted the use of the algorithm-decided grades on August 17th 2020 and all qualifications were awarded the teacher-submitted grade instead (*BBC, 2020*). The public reaction also saw the resignations of Sally Collier, CEO and Chief Regulator and Ofqual, and Jonathan Slater, the most senior civil servant in the Department for Education.

A report published stated no grading bias (*Ofqual, 2020*) but it was found that the algorithm favoured students from more economically privileged backgrounds and meant that students from less economically privileged backgrounds suffered the most (*Smith, 2020*). This led to it being labelled a 'mutant algorithm' by UK Prime Minister Boris Johnson (*Coughlan, 2020*).

METHOD

Data collection and processing

Data collection occurred through using the Twitter for Academic Purposes Application Programming Interface. Using Twitter as a data source here is significant due to the availability of the data, the amount of data and the ability to analyse it in real time. As stated our aim wasn't to conduct a comprehensive analysis of the discourse on social media more broadly, but to examine a specific dataset to critically reflect on computational linguistic methods. Using Tweets as a specific social media dataset comes with many risks and rewards (*Agarwal et al., 2011*). The fact that data can be collected from Twitter in real-time supports the current computational linguistic analysis models that have the functionality to do the same (*Kumar, Morstatter & Liu, 2014*). An important aspect of using Twitter is that its data can be pre-processed before analysis (*Jianqiang, 2015*) and lends itself well to exploratory analysis principles (*Chong, Selvaretnam & Soon, 2014*).

There are some complex ethical considerations when scraping data from Twitter for data analysis. A prominent ethical issue is the fact that although tweets are public(by default), Twitter 'data' isn't provided by users for the purposes of research, and that gaining explicit consent to use their tweets for research is practically infeasible (*Woodfield et al., 2013*). We followed best practices recommended in social media research literature, including not including any screenshots of tweets that may be later identifiable without first gaining consent from the tweet author. Tweets were anonymised to researchers as part of the data cleaning process. This study was approved by our university department's ethics committee.

Data extraction occurred using the Tweepy module in the Python programming language (*Roesslein, 2009*). For the COVID-19 app, the key search criteria for this was tweets containing '@NHSCOVID19app', which is the official Twitter handle for the UK's contact-tracing app, and the related hashtag '#NHSCOVID19app'. The reason for this choice was to ensure that tweets were directly related to experience of the contact-tracing app itself, rather than the NHS Test and Trace system as a whole. Although key parts of the discourse may not be revealed through this search term alone, it provides a starting point for investigating views expressed about the app. In total, 180,281 tweets (1,797,052 words) were taken from 23rd September 2020, the day before the app launched in the UK, to 31st July 2021. After this, tweets were condensed down to remove advertisements from the dataset, resulting in a final dataset of 82,294 tweets over an eleven month period. For the Ofqual algorithm, the search criteria was tweets containing any of the following: 'Ofqual algorithm', 'ofqualalgorithm', 'A level algorithm', 'alevelalgorithm', 'a levels algorithm', 'a-level algorithm' or 'a-levels algorithm'. In total, 18,239 tweets (509,948 words) were taken from 12th August 2020, the day before A Level results were released to students, until 3rd September 2020, after Ofqual's chair appeared at the Education Select Committee. The data was sourced from the United Kingdom and only tweets in English were selected. Therefore, the analysis investigated views expressed in English only.

Data was pseudonymised during extraction, with a unique number generated to refer to each tweet. Stopwords were removed from the dataset using gensim, along with the removal of all long and short URLs and the indication 'RT' (retweet) at the beginning of any tweet. Twitter handles that appeared within tweets were also redacted, using gensim, for anonymity.

Computational linguistic methods

Topic modelling

The topic modelling technique chosen was LDA and the gensim module was chosen to perform this. The existing data was tokenised using gensim's 'simple pre-process' function and bigram and trigram models were created using the 'phrases' function. The process also allowed bigrams to be made and lemmatised using Natural Language Toolkit (*Bird, Klein & Loper, 2009*). The id2word dictionary was inputted and combined with the gensim corpora to create the dictionary specific to this dataset. Each word in the document was given a unique ID. Then, using the dictionary, the corpus, which was a mapping of the word ID and the word frequency, was created (*Rehùøek & Sojka, 2011*). The topics were

then generated and printed using the 'gensim.models.ldamodel.LdaModel' function within gensim.

Sentiment analysis

Two different sentiment analysis approaches were chosen—one that used TextBlob and the other using VADER. This was done to make comparisons between the two systems—especially with the overt claim that VADER accounts for negation within its algorithm. The TextBlob model was programmed with a training set about COVID-19 (*Lamsal, 2021*) to become familiar with COVID-related lexicon using the 'train' and 'test' commands. We did not use this for VADER as it is a pre-trained analyser. For TextBlob, a CSV file containing tweets only was imported into the python library and the command 'blob = TextBlob(sentence)' executed. For VADER, the function was 'sentiment_analyzer_score' and parameters were set so that each tweet could labelled 'positive', 'negative' or 'neutral'—with 0.05 and above being 'positive' and −0.05 and below being 'negative'.

Emotion detection

The emotion detection module run was EmoLex, also in Python. The 'top.emotions' command was deployed, Which exported a CSV table with each tweet's correlation to each emotion—fear, anger, anticipation, trust, surprise, sadness, disgust and joy—declared, along with labelling the dominant emotion in a tweet in a separate column.

Analysis process

To formalise our process, we searched for a framework that would enable us to engage with the methods and critically reflect on our use of them in each of the case studies. However, this search was unsuccessful—while many offered frameworks that were specific to the methods being used, missing elements included how to frame critical reflection within this and ensure its universal relevance to our three chosen methods. As a result of not being able to find a well-defined whole model, we have combined other existing recommendations for best practice in a process that could be replicated by other social media researchers.

The five steps in this process helped guide us through the process of using each of the methods. This is a combination of borrowed best practices and enables critical reflection through the use of Maclean's model. Once the method of analysis is chosen, depending on what is being examined and the aim of the research (*Heitmann et al., 2020*), the steps we followed were:

1. **Set expectations:** record what you hope to find in the discourse from using computational linguistic methods. Setting expectations is advocated by *Post, Visser & Buis (2017)*, who suggest that, by writing down expectations prior to the start of the data collection and analysis, the reflection after this is complete will be much more fruitful.
2. **View as trajectories:** present data chronologically to show which topics are discussed, the sentiment of views expressed or the emotions detected. This is a good place to begin to see patterns and areas of interest in the data. Presenting longitudinal data as a trajectory is advocated by *Howard (2021)* and compliments how trends can be seen quickly trhough real-time data collection (*Alamoodi et al., 2021*).

3. **Human review:** according to similar studies (*González-Ibánez, Muresan & Wacholder, 2011*; *Maier et al., 2018*), it is important to human review a sample of the tweets. This offers us the opportunity to not only classify the tweets according to the categories defined by each tool, but also annotate instances of potential inaccuracy, such as sarcasm or negation. The human review was undertaken by two different reviewers and inter-annotator agreement calculated.
4. **Examine items of interest with context:** whether they are turning points, extreme polarities or suggest they have been questionably categorised, examining these with contextual data, such as knowledge about events that move the public at the time, may help create more meaning from the results, as per the suggestions of *Agarwal et al. (2015)*.
5. **Conduct formal critical reflection:** formally conduct critical reflection using Maclean's weather model. Use the expectations recorded before using the method to measure its success and suitability for analysis on this occasion.

Critical reflection model

After exploring each of the methods using the process, a critical reflection takes place. *Finlay (2008)* defines critical reflection as 'learning through and from experience towards gaining new insights of self and practice'. Critical reflection is not a new concept within NLP, HCI and social media research (*Sengers, McCarthy & Dourish, 2006*) but a greater focus has been placed in the design process of new or developing technology. Within our work, we employ critical reflection to examine how suitable the method has been to investigate the public discourse surrounding digital contact-tracing in the UK by applying a simple four strand critically reflective model outlined by *Maclean (2016)*. Using principles from other models, such as the one by *Gibbs & Unit (1988)*, Maclean provides the following stages:

- Sunshine—what went well?
- Rain—what didn't go well?
- Lightning—what came as a shock or surprise?
- Fog—what wasn't understood or could be a further challenge?

Although initially used for educational practitioner reflection, this model allows for simple yet robust critical reflections that we thought would provide a concise format to present lessons learnt in an accessible form for other social media researchers like us who are not experts in computational linguistics. Maclean's model's focus on aspects that were surprising or shocking make it different from most models and presents the opportunity to carve out future plans and work from the final reflective stage. It is important to note that the critical reflection is for the use of the method itself, rather than how successful the process was in aiding the analysis. For the purpose of demonstrating this process, critical reflections will take place after reviewing the results of the method for both case studies, allowing us to draw comparisons between the two examples offered. When reflecting, we were particularly interested in the speed, clarity and accuracy of the processes and outputs.

RESULTS

Herein, we present the use and reflection of the computational methods, by example of our two chosen case studies, using the aforementioned process. This is organised by the method.

Topic modelling

Set expectations (Step 1)

NHS COVID-19 app

One of the expectations of using these methods for this case study was to see the broad themes associated with the NHS COVID-19 app that were being discussed online. It was anticipated that each topic would be generated with a distinct set of words that would be associated with it to make it clearly defined. We also expected that the lexical items found within each theme would make it easy to label the topics.

Ofqual algorithm

For this case study, the expectation of applying these methods was to, once again, see whether there were any broad themes within this online discourse relating to the Ofqual algorithm. We anticipated a smaller number of latent topics in comparison to the COVID-19 app case study due to the reduced time frame and dataset size. We again believed that the lexical items would clearly indicate the overarching topic labels.

View as trajectories (Step 2)

NHS COVID-19 app

Three latent topics were discovered through gensim LDA. Each topic contained ten key lexical items. These words are presented in descending order of association with the latent topic in Table 1. The number of topics was decided on through manual topic inspection and regeneration, examining the ten key words each time, to ensure minimal lexical item overlap.

Data should be presented as a trajectory. With regard to how the topics presented themselves in the tweets from each month of the research time frame, Fig. 1 details the percentage of tweets relating to each topic per month.

The discovery of these topics through the use of gensim's LDA function provides starting points for further focus. As shown, the most featured word of the most prominent topic is 'isolate'. This foregrounds the importance of the topic of self-isolation as part of the discourse surrounding this specific contact-tracing app, underlining that this disruptive effect the app can have on people's lives is dominating the social media discourse about it too. It is also of note that 'serco' is the most common word associated with topic two, which could show concern for who is responsible for the design and implementation of the app.

With the plotting of each topic's prevalence in the discourse for each month, it can be seen that Topic 1 is the topic that has been detected in tweets most consistently, although tweets were more concerned with Topic 2 at the time of the app's launch. In November 2020, there appears to be a rise in tweets that discuss Topic 3, and this is the same again in April 2021.

Table 1 Ranking of the top 10 lexical items associated with each latent topic.

	Topic 1	Topic 2	Topic 3
1	isolate	serco	download
2	positive	government	protect
3	contact	phone	help
4	work	work	store
5	notification	data	risk
6	code	download	google
7	phone	iphone	apple
8	results	private	play
9	self	phones	love
10	tested	good	available

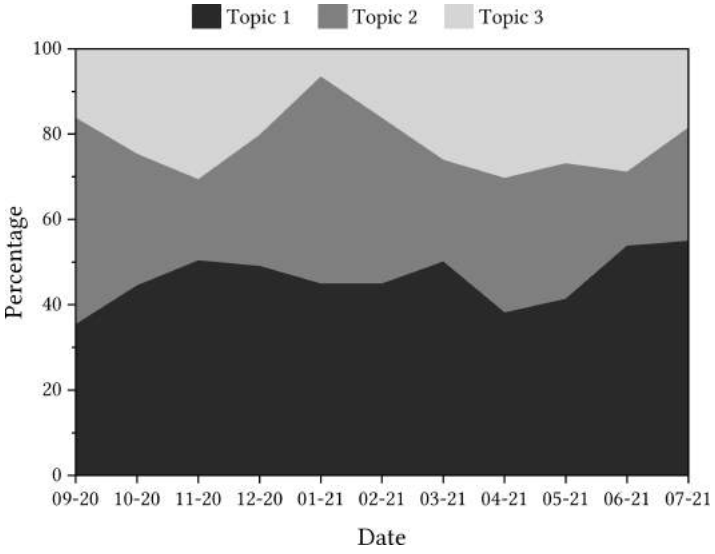


Figure 1 Trajectories of topics detected in tweets containing 'NHSCOVID19app'.
Full-size DOI: 10.7717/peerjcs.1211/fig-1

Ofqual algorithm

Once again using gensim LDA, four latent topics were discovered through gensim LDA, each with ten key lexical items that are associated with that topic, presented in Table 2. Once again, the number of topics was decided on through manual topic inspection and regeneration. Initially, we had three topics, but we concluded a fourth was necessary due one topic containing many generic words and dominating the discourse.

Topic trajectories are presented in Fig. 2.

The one of the most featured words of the most prominent topic is 'government', which may foreground the importance of the role of the UK government in the decision to use and then withdraw the algorithm. 'Flaw' is the most featured word in Topic 2, potentially highlighting there is a significant amount of discussion about whether the algorithm was fit for purpose.

Table 2 Ranking of the top 10 lexical items associated with each latent topic.

	Topic 1	Topic 2	Topic 3	Topic 4
1	level	flaws	exam	students
2	government	statistics	gcses	levels
3	data	father	fiasco	schools
4	algorithms	foresaw	mutants	teachers
5	williamson	punishment	boris	grade
6	like	exams	johnson	school
7	wrong	labour	news	government
8	education	unlawful	blames	downgraded
9	blame	controversial	bbc	based
10	gavin	williamson	ofqual's	teacher

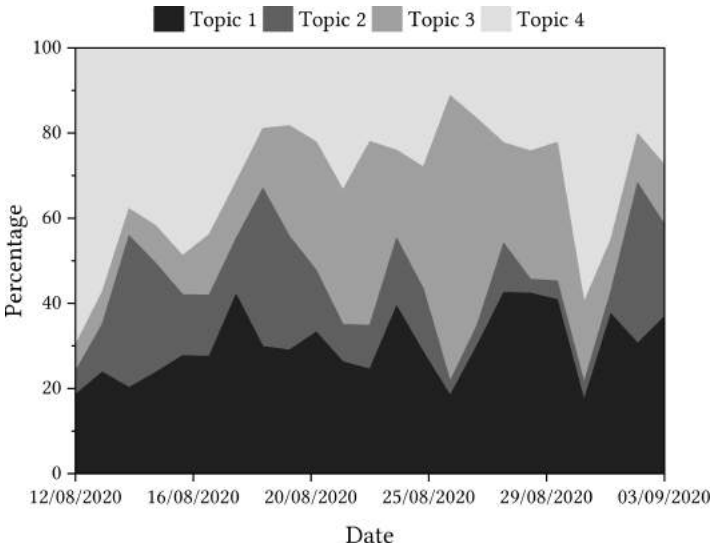


Figure 2 Trajectories of topics detected in tweets relating to the Ofqual A Level algorithm.
Full-size DOI: 10.7717/peerjcs.1211/fig-2

When looking at how the trajectories of the topics fared over the sample period, Topic 4, with words such as 'students', 'schools', 'teachers', etc., dominated the discourse initially. This aligns with the first reporting of the story, where lexical items such as these may have been popular. It might also point to intense focus on the feelings of students, teachers and schools who were affected by the algorithm. The topic did not dominate the discourse again until the end of the month, although, when examining the contextual factors of the algorithm in mainstream media, there is no clear reason why this may have happened.

Human review (Step 3)

NHS COVID-19 app

After this, two blind human reviews were completed. A random sample of 10 tweets per month (110 total) were selected and categorised according to the pre-defined topics that were generated. The reviews found a 57% match between the human reviews and the automated topic labelling. Inter-annotator agreement (measured by Cohen's Kappa) was 0.525, indicating moderate agreement according to *Viera, Garrett et al. (2005)*. This agreement rate may have been due to the broad topics. In this, common errors included labelling of Topic 1 when the automated labelling suggested it would be Topic 2 (and vice-versa).

Ofqual algorithm

For the Ofqual Algorithm dataset, 10 tweets were randomly sampled for each day of the discourse (230 tweets total). The human reviews matched the automatically assigned topic 36% of the time. Inter-annotator agreement was 0.476, again indicating moderate agreement. A reason behind this might have been the separation between two topics that occurred in early stages of the process. The most common error seen here was the human reviewer labelling a tweet associated with Topic 1 as Topic 4 (and vice-versa). This low classification accuracy when comparing the human reviews to the automated topic might indicate that some of the topics are unclear or not fit for purpose.

Examine items of interest with context (Step 4)

NHS COVID-19 app

The next step was to zoom in on items of interest with context. The sharpest monthly increase in a topic's discussion is the rise in tweets discussing Topic 2 in January 2021. When attempting to relate this contextually to the state of the pandemic in the UK, the shift to a more prominent Topic 2 came at the same time as the second nationwide lockdown and, thus, tweets that discuss the government may have increased. This was evidenced when looking at the human review sampled tweets. As shown in Fig. 1, the greater discussion of Topic 2 came at the expense of Topic 3, which was less prominent in tweets in January 2021. This suggests that tweets were less concerned with downloading the app and more concerned with the development of the app by the government and Serco at this point in time. Again, this was seen in the human review sample.

Ofqual algorithm

Discussion around the government and officials within the Department for Education, including Gavin Williamson, as seen in Topic 1, proves popular throughout the sample period according to the data. However, this is very different to how Topic 2, including lexical items such as 'flaw' and 'statistics', is discussed. This topic is popular at the start of the sample period, but declines in proportional popularity until the Educational Select Committee at the beginning of September. Instead, the discourse shifts to the discussion of Topic 3, featuring word such as 'fiasco' and 'mutant'. This topic is seen to gain in popularity after the government u-turn on the use of the algorithm, and is very prevalent in tweets up until the end of August. Upon the announcement of Boris Johnson labelling

the algorithm as 'mutant', this became the overwhelmingly most discussed topic according to the data. The labelling of the event as a 'fiasco' as the story gained in media popularity may have been a contributing factor in the rise of this latent topic, which was reflected in the language seen in the human review sample tweets too.

Critical reflection (Step 5)

Sunshine Upon reflection, the positive aspects of using LDA and topic modelling for both datasets have been the ability to see which lexical items appear frequently with one another, in an attempt to discover latent topics. The gensim tool for topic modelling was also easy to use and compatible with both datasets. Additionally, combining this method with the context of the chosen case studies illuminated topics for potential further analysis and follow-up.

Rain One of the main things that did not work as well with using gensim's LDA topic modelling was that there was limited specific guidelines on how the output (*e.g.*, the lexical items) should be interpreted. As a result, our own interpretation has led us to consider what the topics might be about (although we have refrained from naming the topics in either case study). This means that comparing our findings to those conducting similar studies may be more difficult.

Lightning A surprise that occurred through using LDA was the recurring lexicon that appeared in different topics in both datasets. This foregrounds the importance of context in this process: the same word can have different semantic and pragmatic meanings attached to it when it appears in two different topics (*e.g.*, 'work' as in 'function' or 'paid labour').

Fog An aspect that caused slight confusion through using gensim's LDA is, once again, the interpretation strand. Although interpretation is encouraged after compiling a lexicon associated with topics, it is difficult to pin down how interpretation occurs. Given that LDA is an automated process based on frequencies, a challenge is how a human interprets the results to create meaning.

Sentiment analysis

Set expectations (Step 1)

NHS COVID-19 app

The expectation of using sentiment analysis was to gain an overview of the discourse and see whether trajectories aligned themselves with contextual factors occurring simultaneously. The sentiment analysis results should tell us whether the overall feeling of this discourse is positive, negative or neutral.

Ofqual algorithm

We again wanted to see the general feeling of sentiment detected in the discourse and how co-occurring events may have impacted sentiment. This sentiment may manifest in whether individual parts of the discourse are predicted to be positive, negative or neutral.

View as trajectories (Step 2)

NHS COVID-19 app

From the TextBlob sentiment analysis, Fig. 3 shows that overall sentiment was 0.03 to 0.16, indicating that overall sentiment is slightly above neutral.

After applying TextBlob sentiment analysis, data was presented as a chronological trajectory. The general trend saw positivity detected within tweets rise from September to November 2020, only for tweets to be categorised as more negative in both December 2020 and January 2021. Positive sentiment detected rose again in February and March 2020, dipping slightly in April, but rising again in May and June. Tweets were deemed less positive comparatively in July.

Ofqual algorithm

From the TextBlob sentiment analysis, Fig. 4 shows that overall sentiment ranged from 0.088 to −0.052, indicating that overall sentiment is neutral. However, from the VADER sentiment analysis, Fig. 4 shows that overall sentiment ranged from 0.03 to −0.5, indicating that overall sentiment is negative.

When presenting the results as a trajectory, the general trend saw an increase in negativity on 14th August, with a steady rise in positivity detected in tweets for the next few days. From 17th August, there was an increasing in negativity detected within the tweets until 23rd August, when tweets were more positive. There was a rise in more negative tweets detected in late August but positive sentiment rose again towards the end of the data collection period.

Human review (Step 3)

NHS COVID-19 app

For this human review, 10 tweets per month (110 total) were randomly sampled and classified by two reviewers according to whether they were positive, negative or neutral. The human review score matched the computer assigned sentiment category on 50% of occasions. In terms of individual categories, the reviewers mostly disagreed with the tweets classified as 'positive', with only 22/54 occasions matching the algorithm-generated category. The inter-annotator agreement was 0.62, indicating substantial agreement. Between the reviewers, classifying tweets that the algorithm deemed 'positive' caused the most disagreement, with the reviewers not matching on 16/54 occasions. However, neutral tweets also caused disagreement, with 5/18 classifications not matching between reviewers.

Between the two reviewers, 33.6% of tweets contained negation structures. According to the human reviewers, 56.7% of these tweets containing negation structures were classified incorrectly by TextBlob. This will be explored further in Step 4.

When considering sarcasm, 9% of the tweets reviewed were labelled as sarcastic. 90% of these tweets were classified incorrectly as positive by TextBlob according to the human review. The qualitative examination of these tweets will be expanded on in Step 4.

Ofqual algorithm

Again, 10 tweets per day (230 total) were randomly sampled and classified by two reviewers according to whether they were positive, negative or neutral. The human review score

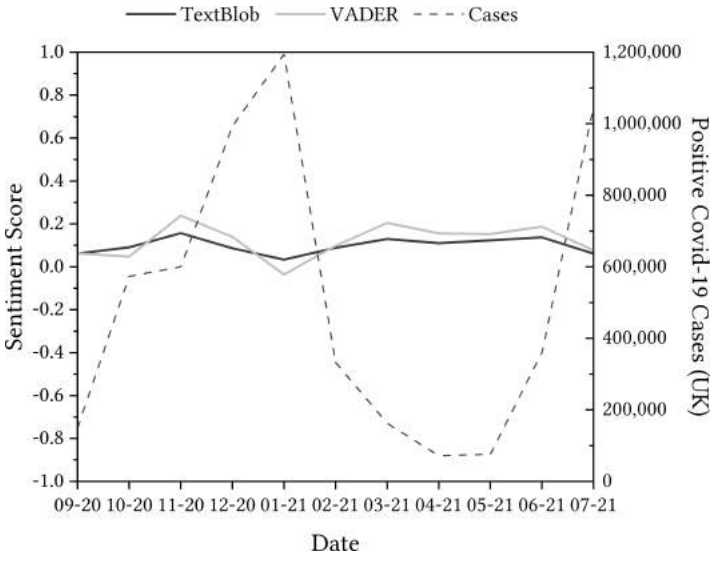


Figure 3 **Evolution of the sentiment of tweets containing 'NHSCOVID19app' using TextBlob and VADER from September 2020 to July 2021.**

Full-size DOI: 10.7717/peerjcs.1211/fig-3

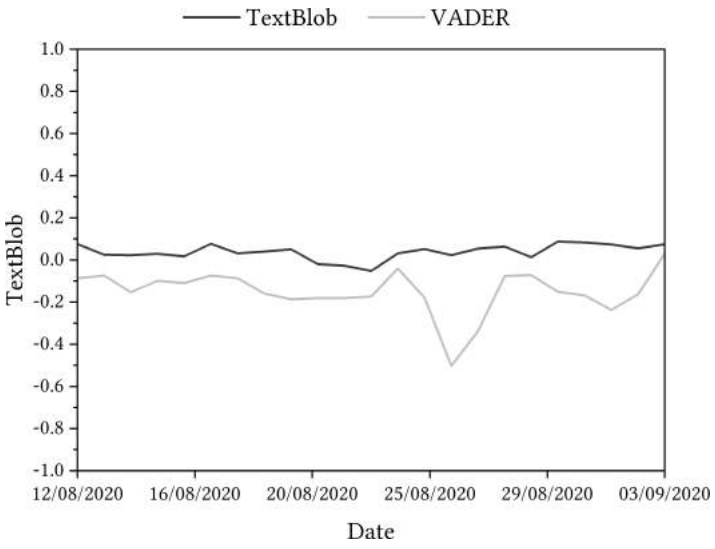


Figure 4 **TextBlob and VADER sentiment analysis of tweets relating to Ofqual A Level algorithm in August and September 2020.**

Full-size DOI: 10.7717/peerjcs.1211/fig-4

matched the computer assigned sentiment category on 40% of occasions. The inter-annotator agreement was 0.547, again indicating moderate agreement. This agreement may have been slightly higher than the topic modelling agreement levels due to the more straightforward classifying process (positive, neutral, negative only). Nevertheless, there were still disagreements between reviewers. Between them, classifying tweets that the

algorithm deemed 'neutral' caused the most disagreement, with the reviewers not matching on 29/72 occasions.

A total of 26% of tweets were found to contain negation structures. According to the human reviewers, 74.6% of these tweets containing negation structures were classified incorrectly by TextBlob. Considering sarcasm, 6% of the tweets reviewed were labelled as sarcastic. 71.4% of these tweets were classified incorrectly as positive by TextBlob according to the human review. A closer look at these tweets takes place in Step 4.

Examine items of interest with context (Step 4)

NHS COVID-19 app

Comparing sentiment detected in tweets relating to the app to the wider context of the UK's history with the pandemic was the next step in the process. For this, we undertook an examination of positive COVID-19 cases and the sentiment detected in tweets, depicted in Fig. 3 as they showed similar inverted trends—sentiment for the app is more negative as the number of cases rise in January, and sentiment for the app is more positive as the cases lower in the spring months. Note that we don't want to suggest causation here, but to highlight the importance of wider contextual data which are vital to interpret sentiment analysis results at the time at which the data was recorded.

Additionally, from the human review sample selected, there were a number of tweets which had been classed as positive (1.0). Of these tweets, there were some of note that had this classification but, upon inspection, it is likely that other human classifiers would agree with us in making an alternative decision, and decide that they were expressing mainly negative sentiment instead. An aspect of language that these tweets have in common is the inclusion of negation within their syntactical structures (*e.g.*, "I don't"; "can't say"; "not to be")', together with words that taken without negation could be interpreted by an algorithm as very positive (*e.g.*, "trust", "best", "impressed", "proud"). Tweets that could be interpreted as sarcastic were also of note here for similar seemingly inaccurate detection reasons. Several tweets in the data sets were all categorised as tweets that had strong positive sentiment detected when classified using the TextBlob module but were categorised as negative when examining with context.

The seemingly incorrect categorisation of these tweets could mean that the reliability of the TextBlob sentiment analysis tool may be questioned. Therefore, we deployed another sentiment analysis module for further investigation. Due to the fact that the VADER sentiment analysis module states that it accounts for negation, it was a logical next step to see whether there was a difference in the categorisation of the dataset, but also with these focused tweets more specifically.

When comparing the two, as shown in Fig. 3 the general trend of the data when comparing the VADER sentiment analysis to the TextBlob sentiment analysis is similar, but varies in two different places: October 2020 and May 2021. There are opportunities to investigate these differences in sentiment polarity detection through a critical lens, but it could be said that sentiment analysis alone might offer limited capability to do this without the inclusion of a qualitative method in conjunction. That being said, the comparison

could still be argued as noteworthy: the polarity of VADER appears to be more extreme than TextBlob—seeing deeper rises and falls in sentiment detected in tweets.

With regard to the tweets that were categorised as positive by TextBlob, despite finding negation in the tweets, these were classified as −0.4023, 0.6369, 0.4767 and 0.2924. This suggests that, while there is some improvement in detecting negation within VADER, both VADER and TextBlob and other easily accessible sentiment analysis models may still benefit from further improvements of this aspect of language to improve the accuracy of their tools; but perhaps more importantly, users of these methods like us would benefit from greater transparency regarding these limitations of the models.

For tweets that may be sarcastic, the sentiment detected was still positive with the VADER module, with values of four particular tweets being 0.8316, 0.7622, 0.6767 and 0.6958 (rather than 1.0). Again, this suggests that, even though the VADER system was potentially able to detect sarcasm better than TextBlob, both systems may benefit from further focus on this. Additionally, users would benefit from more transparency regarding these important limitations.

Ofqual algorithm

To interpret the results as best as possible, they were examined with key dates in the chronology of the Ofqual algorithm, as per the next step in the process. For example, a negative change in sentiment was detected on August 14th, the day after the results were shared with students. A rise in positive sentiment detected on 17th August came on the same day as the government announced that the algorithm-calculated grades would be replaced with teacher assessed grades. The sharp negative change in sentiment on 26th August came on the same day as UK Prime Minister Boris Johnson told students that their results had been affected by a 'mutant algorithm'. A rise in positive sentiment on 3rd September came the day after Ofqual Chair, Roger Taylor, apologised to students when appearing at the Educational Select Committee at the House of Commons.

Similarly to the COVID-19 app case study, there were instances where the algorithm classification involving negation and sarcasm could be seen as incorrect. Within the dataset, tweets that were categorised as positive did include negation, pairing 'not' with 'believe' and 'no' with 'accident', for example. Once again, there were many tweets that could be deemed to be sarcastic that were classed as positive by TextBlob. With these tweets, hyperbolic adjectives such as 'sophisticated', 'great' and 'flawless' all seen.

As a result, the VADER module was deployed. With the inclusion of the VADER sentiment data, as displayed in 4, overall sentiment ranged from 0.03 to −0.5, indicating that sentiment is negative. On 23rd August, there was an abrupt change in sentiment detected in tweets, as the data in Fig. 4 shows—sentiment detected increased in positivity. This is perhaps more pronounced when looking at the shift in sentiment that follows in the coming days—a change in sentiment to −0.5 on 26th August. Tweets were deemed more positive comparatively at the beginning of September.

When revisiting the tweets in the sample that included negation but were potentially incorrectly classified, new VADER scores were 0.8372, 0.6546, −0.1655 and 0.4019, showing a similar result to the COVID-19 app's VADER scores for negation specifically. For tweets

deemed sarcastic, the VADER scores were 0.9739, 0.3612, 0.1548 and 0.9081. Once again, this replicates the findings from the COVID-19 app case study.

Critical reflection (Step 5)

Sunshine A positive aspect of using this method is the ability to analyse a large amount of data within a seemingly short amount of time, with a straightforward coding process. This was seen in both case studies. Another aspect that worked well was how simple the two methods were to compare after deciding to integrate VADER into the sentiment analysis process. Sentiment scores over a larger period time, such as shown in Figs. 3 and 4 provide an instant overview of the trend of sentiment (*i.e.,* 'going up and down') in the dataset over time. This lends itself to identify 'turning points' for further investigation, potentially through in-depth qualitative methods.

Rain An aspect of investigation that did not work as well was the fact that negation and sarcasm are still unsolved challenges in sentiment analysis and may therefore decrease the accuracy of the techniques and the robustness of the findings. Individual 'sentiment scores' are hard to interpret and feel meaningless on their own.

Lightning A surprising occurrence when undertaking sentiment analysis using TextBlob and VADER was the variation between outcomes seen, as it poses questions as to which is the more reliable tool to use for conducting a sentiment analysis at this scale. Additionally, for the Ofqual algorithm, the variation seen in the data—where sentiment changes dramatically—was surprising, particularly when contextually there was little that could be illuminated with some of the changes (like 24th August).

Fog A confusing part of using sentiment analysis comes from the idea of interpreting the data. When a sentiment score is generated, there is limited guidance on what that score means and how it can illuminate further insights into the status of the contact-tracing discourse in the UK. Although these scores can be used as a starting point, presenting them without context may leave a lack of clarity. For the Ofqual algorithm, there is limited guidance on what a sentiment score means and how it can illuminate further insights into the status of the discourse, leading to a lack of clarity. For example, what is the difference between a score of −0.18091 and −0.1815, as seen on 21st and 22nd August in the VADER classification for the Ofqual algorithm?

Emotion detection

Set expectations (Step 1)

NHS COVID-19 app

The expectation of using emotion detection was that using this method would allow an insight into how people felt about the app and whether there were any shared or common emotions expressed. We also expected the results to tell us which emotions were more prevalent at certain times within the longitudinal discourse.

Ofqual algorithm

Our expectation was to see which emotions were the most common in the discourse and how the emotions detected may change over time. With this being a shorter time period,

we expected to see perhaps not as many emotions detected or possibly not as much gradual change in emotions predicted.

View as trajectories (Step 2)

NHS COVID-19 app

For the next step in the process, the data presented in the trajectory displayed in Fig. 5 shows that 'trust' was the emotion most detected in tweets relating to the app, followed by 'fear' and then 'anticipation'. By examining the percentage of tweets that detected each emotion, the rise in tweets related to 'fear' at the end of 2020 could be deemed to be of interest, as could the rise in tweets related to 'anticipation' in the spring of 2021.

Ofqual algorithm

Once again, as shown in Fig. 6, 'trust' was the emotion most often detected, followed by 'fear'. Looking at the trajectory of emotions detected in the sample period, there are key dates that appear to have greater spikes in 'fear', and thus a downturn in 'trust' detected, including the days following the u-turn announcement, the statement about 'mutant algorithm' and 30th August. However, once again, it remains unclear whether tweets relating to 'trust' are indeed positive indicators of trust or indicators of mistrust.

Human review (Step 3)

NHS COVID-19 app

For the COVID-19 app, ten tweets per month (110 total) were randomly sampled to be reviewed. The categories to be assigned were 'trust', 'fear', 'anticipation', 'anger', 'surprise', 'sadness', 'disgust', 'joy' and 'no emotion'. Reviewers matched the EmoLex assigned category on 25% of occasions. The inter-rater reliability was 0.44, indicating moderate agreement. These agreement levels may be lower than the topic and sentiment agreement levels due to the range of emotions available to classify with. Within this, between the reviewers, classifying tweets that the algorithm deemed as 'anger' caused the most disagreement, with the reviewers not matching on 5/11 occasions. Reviewers categorised these tweets as 'fear' or 'disgust' instead.

Ofqual algorithm

For the Ofqual discourse, ten tweets per day (230 total) were randomly sampled to be reviewed. Once again, the categories to be assigned were 'trust', 'fear', 'anticipation', 'anger', 'surprise', 'sadness', 'disgust', 'joy' and 'no emotion'. Reviewers matched the EmoLex assigned category on 37% of occasions. The inter-rater reliability was 0.481, again indicating moderate agreement. Once again, these agreement levels may be explained by the range of emotion options available. Tweets reviewed in this process will form the examples of the following section. Between the reviewers, classifying tweets that the algorithm deemed as 'fear' caused the most disagreement, with the reviewers not matching on 17/60 occasions. Instead, reviewers categorised these tweets as 'anger', which is the reversal of what was seen in the COVID-19 app dataset.

Examine items of interest with context (Step 4)

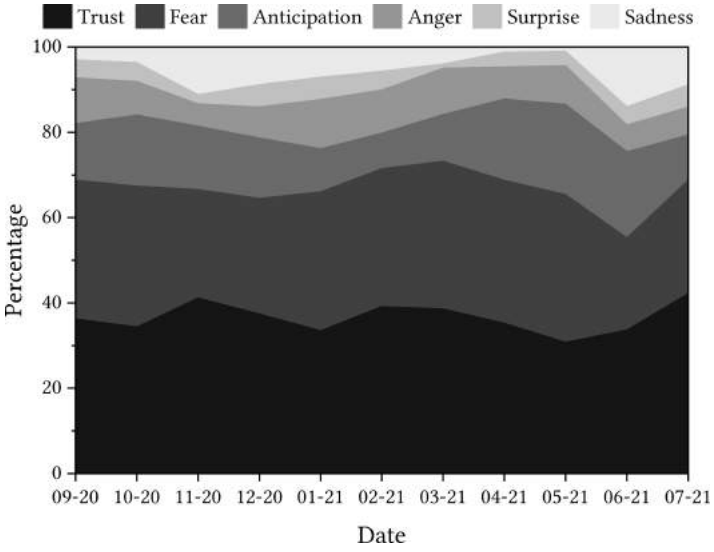


Figure 5 **Emotions detected in tweets relating to 'NHSCOVID19app'.**
Full-size DOI: 10.7717/peerjcs.1211/fig-5

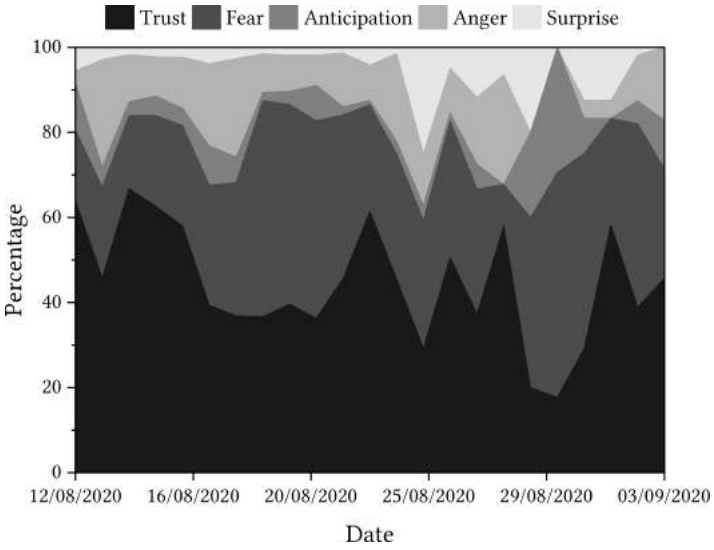


Figure 6 **Emotions detected in tweets relating to the Ofqual A Level algorithm.**
Full-size DOI: 10.7717/peerjcs.1211/fig-6

NHS COVID-19 app

When zooming in, the consistency of 'trust' being an emotion that is detected within tweets relating to the app means that it is a prominent emotion of discussion. With the premise that 'trust' and 'fear' are separate emotions, it might be assumed that the tweets relating to 'trust' are indicative of the person trusting the app, rather than not trusting it. This thought guided us through the next part of the process.

Some tweets had been classified as containing the emotion 'trust' within them; again, we would have categorised these differently due to the opposition to trust noted within each

tweet. The tweets included the word 'trust', which had been negated using the contracted modal verb 'wouldn't'. This may have not been detected by the EmoLex module and, instead, the word 'trust' superseded any other vocabulary as the classifier made its decision. This is similar to the negation issue using TextBlob for sentiment analysis. However, there are other tweets that have been noted as having 'trust' detected within them that do not contain the word 'trust'. Although it may be possible to argue that the words 'faith' and 'encourage' could hold some similarity in meaning with the word 'trust', there is no explicit mention of 'trust' within these tweets and a human might categorise them as expressing a lack of trust in the app, again as negations are present (*e.g.*, ''not downloading'', ''haven't got faith''). This further demonstrates a possible potential limitation of using tools such as EmoLex for emotion detection.

Ofqual algorithm

One potentially interesting emotion to explore is 'anticipation'. While there were some tweets throughout the discourse that were categorised as expressing the emotion 'anticipation', this is a relatively low amount of tweets until 30th August, when 11.63% of tweets were classified in this category, rising sharply from 4.17% the day before and no tweets detected two days before. Once again, 30th August does not appear to be a significant date in the timeline of events relating to the Ofqual algorithm, and so the potential rationale for this increase is not immediately clear.

'Anger' and 'surprise' are the two other emotions detected in the dataset. 'Anger' appears to be more prevalent in the discourse at the start of the sample period, particularly on the day that results were released and the day of the government u-turn. With 'surprise', however, this emotion is detected in relatively few tweets at the start of the discourse, which seems to change on 26th August—the date that Johnson made a statement on the 'mutant algorithm'. This 'surprise' might be attributed to the use of the word 'mutant', or the general unexpected nature of the statement. In what might seem an apparent consequence of this statement, 'surprise' is generally detected more frequently that previously in the majority of the remaining days in the sample period.

As with the COVID-19 app case study, there are questions as to whether the emotion 'trust' indicates the direction of trust. For example, tweets containing 'criticised', 'unfair', 'unequal' and 'failure' were all attributed to 'trust', despite being potentially more indicative of distrust instead. Additionally, when presented with a tweet that contained vocabulary such as 'great' and 'similar' (used in a literal and not sarcastic context), there was limited opportunity to classify this as 'happy' or 'supportive'. For context, EmoLex had classified this tweet as 'anticipation'. Going forward, this might be a consideration for developing expectations when working with emotion detection algorithms, which feeds into the process.

Critical reflection (Step 5)

Sunshine One of the aspects that worked well here was the speed of automated detection with a large dataset and that each tweet was able to be classified in some way.

Rain As with sentiment analysis, the accuracy of the detection may have been an issue again when deploying the EmoLex emotion detection module. Additionally, insights into

the discourse are presented without context—something that would have been helpful in the analytical process. With limited context, the emotions could be seen as arbitrary.

Lightning A surprising element of using emotion detection was the inclusion of 'positive' and 'negative' within the initial emotion set in EmoLex. This might have meant that vital information may have been missed as the tweets may have been closely aligned with other emotional categories, yet this has not been included in the results.

Fog Something that could have been clearer when using this tool was the distinction between emotions. For example, should tweets associated with trust only feature those that actively support this emotion, rather than oppose it? As seen in further exploration, there were tweets categorised as containing the emotion 'trust' that may have been categorised to be in opposition to trusting the app if classified by a human. If this had been clearer before using the tool, the picture of the discourse may have been more accurate.

DISCUSSION

The following section will take the form of two main parts. The first part discusses each step of the process. The second part looks at the implications that our research has on current and future research in the field of social media research.

The analysis process

Set expectations (Step 1)

Through the setting of expectations, we were able to make clear what we aimed to discover through using the methods on each case study. These were particularly useful to refer to when undertaking the critical reflection as a measure of what we anticipated might be unearthed during the analysis of results, which supports the advocacy of *Post, Visser & Buis (2017)* on this topic. This will be revisited in the discussion of the critical reflection later.

View as trajectories (Step 2)

The analysis and trajectories, presented by the method used, will now be discussed in detail. Displaying the data as chronological trajectories was helpful in identifying changes in the discourse with regard to differing topics and emotions, and how sentiment fluctuated. Even though *Howard (2021)* advocated that this is useful for longitudinal datasets and discourses, this was seen as beneficial to both the longitudinal COVID-19 app discourse and the snapshot A Level algorithm discourse.

Topic model analysis, as a whole, used adjacent words well to describe latent topics present within the text, which is similar to how *Hu, Chancellor & De Choudhury (2019)* used LDA to aid their investigation into homelessness discourses online too. Additionally, gensim's ease of use and compatibility with the Twitter dataset meant that it was a good fit to begin the analytical process with this particular social media platform, much like *Hidayatullah et al. (2019)* and their use of gensim. Seeing the topics presented as trajectories allowed us to use them as a springboard for further investigation, like *Song et al. (2019)*. However, the process for deciding how many topics were chosen for each discourse could have been improved and may have impacted some of the errors found later on with classification. For example, in the Ofqual discourse, we manually decided to expand from

three topics to four (*Nguyen et al., 2020*), but instead we could have used a measure for topic coherence, such as Hierarchical Dirichlet Process (*Teh et al., 2004*), to arrive at an optimal number.

For sentiment analysis, the trajectories showed a large amount of data that was able to be analysed through both TextBlob and VADER, which was the starting point in gaining some insights. The comparison between the two systems, as another was required, was easy to facilitate too, much like how *Pak & Paroubek (2010)* compared classifiers. However, these trajectories could only tell us a small amount of information to begin with, and it was clear that more interpretation and analysis was necessary. Nevertheless, it was good to know where to focus our attention moving into the next steps (where items of interest are examined with context).

Reflections regarding emotion detection are similar to those concerned with sentiment analysis. The ease of deployment and the fact that categorisation could occur for every tweet was an advantage of using this method, which meant that displaying this as a trajectory aided our identification of items of interest. This aided the showcase frequently detected emotions, which links to earlier strengths of the method discussed in the work of *Mathur, Kubde & Vaidya (2020)* and *Aribowo & Khomsah (2021)*.

Human review (Step 3)

Undertaking the human review meant that the algorithm-generated results could be compared against human classification. The results found a low percentage of classification accuracy when comparing the human and algorithm-generated scores, ranging from 25% to 50% match. This, at surface level, indicates that there are potential errors in the classification, which may be from either the human or algorithm score.

For topic modelling, the main finding was that the topics may not have been accurate, particularly for the Ofqual discourse, due to the low percentage accuracy (57% for the NHS discourse and 35% for the Ofqual discourse). As we followed the principles of *Nguyen et al. (2020)*, we thought that having fewer topics would have led to fewer inaccuracies. However, it may have been more beneficial to have refined the topics to create smaller, more-defined themes, like *Sengupta (2019)*. This was particularly true for the Ofqual discourse.

For sentiment analysis, this process not only allowed us to see the accuracy of sentiment detection, but also allowed us to see the proportion of sarcastic and negated tweets that were classified seemingly incorrectly. The figures of 56.7% (NHS) and 74.6% (Ofqual) of negated tweets and 90% (NHS) and 71.4% (Ofqual) of sarcastic tweets being classified seemingly incorrectly by the automated classification indicate that these issues were prevalent, much like the work drawn upon by *Gupta & Joshi (2021)* and *González-Ibánez, Muresan & Wacholder (2011)*. This was a similar process for emotion detection, instead highlighting the potentially incorrect classification of 'trust'. Both of these will be expanded upon in Steps 4 and 5.

The potential questionable output from the 'off-the-shelf' tools had implications for our analytical process. Agreement between human and algorithm classifications was better for the NHS App on two occasions (57% *vs* 36% for topic modelling, 50% *vs* 40% for sentiment analysis) and better for the Ofqual algorithm on one occasion (37% *vs* 25%

for emotion detection). Regarding subjectivity, the discrepancies may question what is understood to the 'ground truth' in these scenarios. More specifically, occasions where the human reviewers and the algorithm-generated score or classification were in disagreement could have been due to two reasons: ambiguity in the labels or the idea that some text holds meaning on multiple levels (*Fujioka et al., 2019*). These concepts are usually difficult for computational linguistics to detect and interpret. This meant that we proceeded with Step 4 with certain groups of tweets as immediate focal points, especially when concerned with changes to potential trajectories or the expectations previously set. When reflecting, this meant that the latter steps in the process may have been influenced by these low agreement rates.

The human review aspect could benefit from further development. As reported, there were occasions where the human reviewers did not agree on classification, with inter-annotator agreements usually being moderate. This was particularly prevalent when classifying neutral tweets in the sentiment analysis and classifying tweets relating to anger and fear in the emotion detection. Seeing that humans may disagree to a significant extent about posts belonging to the same classification, or a post's polarity, has implications for how annotation works in text mining more broadly. For instance, emotion categories that may overlap, such as fear and anger, may cause discrepancies (*De Silva et al., 2018*; *Leung et al., 2021*). Also, some reviewers classify tweets including 'positive' and 'negative' elements as 'neutral', whereas others may find the best fit, exemplifying the complexities of agreeing on sentiment (*Villena-Román & Garcıa-Morera, 2013*). Supporting reviewers with examples of lexicon, or more complex emotional categories (*Jiang, Brubaker & Fiesler, 2017*) that could be included in certain categories through training may mitigate such errors and encourage objective discussion.

Additionally, although we were able to see where things may have been classified potentially incorrectly, a formal approach was not undertaken here. Fortunately, there are studies within social media studies that use computational linguistic methods that have included a human review aspect, whether formalised or in a more informal approach, such as *van Atteveldt, van der Velden & Boukes (2021)* and their use of human-trained coding that outperformed machine learning approaches. The inclusion of a more formal qualitative aspect here would aid the analysis and categorisation in the human review stage. As a result, a formal qualitative method could be used, such as critical discourse analysis, which would aid classification through appropriate supporting theoretical frameworks.

Examine items of interest with context (Step 4)

Using contextual information to illuminate the analysis of the discourses proved rewarding and challenging. Using contextual information, at the suggestion of *Agarwal et al. (2015)*, meant that more meaning was created from the results. This was particularly pertinent in the Ofqual algorithm case study due to the more focused timeline. Zooming in closely on these items of interest that were identified in the Step 3 also provided further insight into the potential accuracy of the methods used. For topic modelling in this stage, the limitations and challenges recognised were collectively aligned with how to interpret the

results to produce more meaningful insights. This may be seen to build on the previous work by *Maier et al. (2018)*.

For sentiment analysis, with the limited guidance on what the scores reflect or mean, this adds little to the earlier findings of the study by *Pokharel (2020)* on the Twitter response to COVID-19 in India due to the limited distinction in results when using sentiment analysis alone. Therefore, our findings here are more in agreement with *Sivalakshmi et al. (2021)*, with their assertion that TextBlob being unable to read tokenized special characters may have hindered its performance and suitability here, and *Chauhan, Bansal & Goel (2018)*, who believe VADER's sensitivity to social media formats make it the more appropriate choice. Additionally, with the limited guidance on what the scores reflect or mean, this adds to the use case of sentiment analysis. This reinforces the points made in studies such as *Jiang, Brubaker & Fiesler (2017)*. The fact that this was only illuminated through our critical reflection process shows that there is more work to be done in understanding the limits of sentiment analysis, echoing the work of *Stine (2019)*.

When examining with context in the emotion detection output trajectories, it raised issues as to whether whether trust and distrust should exist within the same emotional 'trust' umbrella in EmoLex, echoing the need for more distinct emotional categories (*Jiang, Brubaker & Fiesler, 2017*). The reflections on EmoLex in this contribution are, therefore, more in line with the improvements suggested by *Balakrishnan et al. (2019)*, *Balakrishnan & Kaur (2019)* and *Fast, Chen & Bernstein (2016)* that concern the intricacies of emotions.

Critical reflection (Step 5)

The use of Maclean's critical reflection weather model provided a format for the process of discovering advantages and limitations of the methods by providing specific focal points. The advantages it highlighted were that the examined computational methods were easy to implement and provided a starting point for further investigation into the discourse. Exploring how views expressed altered through the course of the app's existence, and viewing these as linear trajectories, aided the investigation. The limitations of the methods that the reflection highlighted include the instances of diverging interpretations when examining linguistic features in context; this means that the accuracy of the methods may be brought into question. This was the case for sarcasm and negation in sentiment analysis, the naming and accuracy of topics in LDA (introducing biases as mentioned by *Saura, Ribeiro-Soriano & Saldaña, 2022*) and the distinction between emotions in emotion detection. It is possible that a different model may have been able to mitigate these potential inaccuracies, such as how VADER outperformed TextBlob in the sentiment detection of negation; however, these models may not be as accessible to social media researchers.

Implications for research

This study has several implications for social media researchers wishing to investigate the state of online discourses using descriptive and predictive computational linguistic methods. Firstly, the examination of these methods through a critically reflective lens echoes some of the existing social media analysis critiques of topic modelling, sentiment analysis and emotion detection algorithms.

On a related note, the process we have formed, by combining other models and ideas for existing best practice, may be developed into something more formal and robust: it has the potential to act as a foundation for those unfamiliar with these methods but may wish to use them for exploring public discourses. Social media researchers using this approach may apply these computational linguistic methods to analyse large social media datasets with confidence. It means researchers can realistically manage expectations through the logging of what the researcher wishes to achieve so this can be revisited in the critical reflection aspect, as well as doing more than looking at the outputs of the predictive algorithms by displaying trajectories and examining with context.

When considering future work, this should investigate the ways in which computational linguistic methods may be used more effectively by combining them with qualitative research methods, such as corpus linguistics or discourse analysis. This may help to mitigate some of the limitations of computationally descriptive and predictive analytical approaches that were found during the process of critical reflection in this article. The strong emphasis on context, particularly in critical discourse analysis, means that how views are expressed and related to the events occurring at the time can play a more prevalent role in the analysis of public discourses.

CONCLUSION

In applying popular computational linguistic methods to Twitter case studies, we demonstrate the potential use of a process that combines existing best practice and critical reflections on the use of the methods for social media researchers, like us, who are not experts in computational linguistics. Our findings show that, by following this analysis process, we were able to do more with the results generated by the algorithms through the examination of contextual factors and chronological trajectories post-human review. Our critical analysis and reflection shows that the classification of tweets as expressing a certain sentiment or emotion is not always accurate, with negation and sarcasm standing out as linguistic phenomena that algorithms struggle with. This was reinforced by the issues highlighted with both human-algorithm agreement and inter-annotator agreement. The threads of trust and fear in both datasets, which were detected through emotion modelling algorithms, showed that there is deeper exploration needed to ascertain the cause of these emotions and the degree to which they are felt; this is possibly where a more discursive sociolinguistic method may be helpful. Finally, the use of Maclean's reflective toolkit enabled concise evaluations of the performance and suitability of each of the methods and raised potential concerns with interpretation and lacking transparency regarding limitations such as as dealing with negation and sarcasm. Overall, the critical reflections found from following this process should begin to support the confident and realistic use of computational linguistic methods by social media researchers. NLP approaches should be applied with care and a critical reflection on their output. While the analytical process that includes current existing best practices is a good start, more formal qualitative interpretation of results could be necessary.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

All authors are supported by the UKRI Trustworthy Autonomous Systems Hub (UKRI Grant No. EP/V00784X/1). Dan Heaton is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (UKRI Grant No. EP/S023305/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
UKRI Trustworthy Autonomous Systems Hub: EP/V00784X/1.
Horizon Centre for Doctoral Training at the University of Nottingham: EP/S023305/1.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Dan Heaton conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jeremie Clos conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Elena Nichele analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Joel Fischer conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data and code are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1211#supplemental-information.

REFERENCES

Abbasi K. 2021. *Covid-19: the UK's political gamble that bodes ill for health and the health service.* London, UK: British Medical Journal Publishing Group.

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ. 2011. Sentiment analysis of Twitter data. In: *Proceedings of the workshop on language in social media (LSM 2011).* 30–38.

Agarwal B, Mittal N, Bansal P, Garg S. 2015. Sentiment analysis using common-sense and context information. *Computational Intelligence and Neuroscience* **2015**:715–730.

Alamoodi A, Zaidan BB, Zaidan AA, Albahri OS, Mohammed K, Malik RQ, Almahdi E, Chyad M, Tareq Z, Albahri AS. 2021. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. *Expert Systems with Applications* **167**:114–155 DOI 10.1016/j.eswa.2020.114155.

Arianto BW, Anuraga G. 2020. Topic modeling for Twitter users regarding the "Ruangguru" application. *Jurnal ILMU DASAR* **21(2)**:149–154 DOI 10.19184/jid.v21i2.17112.

Aribowo AS, Khomsah S. 2021. Implementation of text mining for emotion detection using the lexicon method (Case Study: tweets about Covid-19). *Telematika* **18(1)**:49–60 DOI 10.31315/telematika.v18i1.4341.

Arzt S, Poller A, Vallejo G. 2021. Tracing contacts with mobile phones to curb the pandemic: topics and stances in peoples online comments about the official German contact-tracing app. In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, 1–7.

Balakrishnan V, Kaur W. 2019. String-based multinomial Naïve bayes for emotion detection among Facebook diabetes community. *Procedia Computer Science* **159**:30–37 DOI 10.1016/j.procs.2019.09.157.

Balakrishnan V, Martin MC, Kaur W, Javed A, Javed A. 2019. A comparative analysis of detection mechanisms for emotion detection. *Journal of Physics: Conference Series* **1339**:12–16 DOI 10.1088/1742-6596/1339/1/012016.

BBC. 2020. A-levels and GCSEs: U-turn as teacher estimates to be used for exam results. *BBC News. Available at https://www.bbc.com/news/uk-53810655*.

Bird S, Klein E, Loper E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol, CA, USA: O'Reilly Media, Inc.

Bollen J, Mao H, Pepe A. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the International AAAI Conference on Web and Social Media* **5(1)**:450–453.

Chaithra VD. 2019. Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments. *International Journal of Electrical and Computer Engineering* **9(5)**:4452–4459 DOI 10.11591/ijece.v9i5.pp4452-4459.

Chauhan VK, Bansal A, Goel A. 2018. Twitter sentiment analysis using vader. *International Journal of Advance Research, Ideas and Innovations in Technology* **4**:485–489.

Chong WY, Selvaretnam B, Soon L. 2014. Natural language processing for sentiment analysis: an exploratory analysis on tweets. In: *2014 4th international conference on artificial intelligence with applications in engineering and technology*, 212–217 DOI 10.1109/icaiet.2014.43.

Coughlan S. 2020. A-levels and GCSEs: boris Johnson blames 'mutant algorithm' for exam fiasco. *BBC News. Available at https://www.bbc.com/news/education-53923279*.

De Silva D, Ranasinghe W, Bandaragoda T, Adikari A, Mills N, Iddamalgoda L, Alahakoon D, Lawrentschuk N, Persad R, Osipov E. 2018. Machine learning to

support social media empowered patients in cancer care and cancer treatment decisions. *PLOS ONE* **13(10)**:e0205855 DOI 10.1371/journal.pone.0205855.

Dowthwaite L, Fischer J, Vallejos EP, Portillo V, Nichele E, Goulden M, McAuley D. 2021. Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study. *Journal of Medical Internet Research* **23(9)**:e29085 DOI 10.2196/29085.

Fast E, Chen B, Bernstein MS. 2016. Empath: understanding topic signals in large-scale text. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, 4647–4657.

Finlay L. 2008. Reflecting on 'Reflective practice'. *Practice-based Professional Learning*. Milton Keynes, UK: The Open University.

Fujioka T, Bertero D, Homma T, Nagamatsu K. 2019. Addressing ambiguity of emotion labels through meta-learning. ArXiv preprint. arXiv:1911.02216.

Gibbs G, Unit FE. 1988. *Learning by doing: a guide to teaching and learning methods*. Oxford, UK: Further Education Unit.

González-Ibánez R, Muresan S, Wacholder N. 2011. Identifying sarcasm in Twitter: a closer look. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 581–586.

Gupta I, Joshi N. 2021. A review on negation role in Twitter sentiment analysis. *International Journal of Healthcare Information Systems and Informatics* **16(4)**:1–19 DOI 10.4018/ijhisi.20211001oa06.

Heitmann M, Siebert C, Hartmann J, Schamp C. 2020. More than a feeling: benchmarks for sentiment analysis accuracy. *Communication & Computational Methods eJournal* DOI 10.2139/ssrn.3489963.

Hidayatullah AF, Hidayatullah AF, Aditya SK, Karimah , Gardini ST. 2019. Topic modeling of weather and climate condition on Twitter using latent dirichlet allocation (LDA). *IOP Conference Series: Materials Science and Engineering* **482**:12–33 DOI 10.1088/1757-899x/482/1/012033.

Howard AL. 2021. A guide to visualizing trajectories of change with confidence bands and raw data. *Advances in Methods and Practices in Psychological Science* **4(4)**:25152459211047228.

Hu A, Chancellor S, De Choudhury M. 2019. Characterizing homelessness discourse on social media. In: *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, 1–6 DOI 10.1145/3290607.3313057.

Jiang JA, Brubaker JR, Fiesler C. 2017. Understanding diverse interpretations of animated GIFs. In: *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, Chi Ea '17*. New York, NY, USA: Association for Computing Machinery, 1726–1732 DOI 10.1145/3027063.3053139.

Jianqiang Z. 2015. Pre-processing boosting Twitter sentiment analysis? In: *2015 IEEE international conference on smart City/SocialCom/SustainCom (SmartCity)*. Piscataway: IEEE, 748–753 DOI 10.1109/SmartCity.2015.158.

Kelly A. 2021. A tale of two algorithms: the appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal* **47(3)**:725–471.

Kent C. 2020. A comedy of errors: the UK's contact-tracing apps. New York, NY, USA: Medical Device Network.

Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. 2020. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health* **5(8)**:e452–e459.

Kumar S, Morstatter F, Liu H. 2014. *Twitter data analytics*. New York, NY, USA: Springer.

Lamsal R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* **51(5)**:2790–2804 DOI 10.1007/s10489-020-02029-z.

Leung YW, Wouterloot E, Adikari A, Hirst G, De Silva D, Wong J, Bender JL, Gancarz M, Gratzer D, Alahakoon D. 2021. Natural language processing–based virtual co-facilitator for online cancer support groups: protocol for an algorithm development and validation study. *JMIR Research Protocols* **10(1)**:e21453 DOI 10.2196/21453.

Liu B. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing* **2(2010)**:627–666.

Liu B, Zhang L. 2012. A survey of opinion mining and sentiment analysis. In: *Mining Text Data*. 415–463 DOI 10.1007/978-1-4614-3223-4_13.

Loria S. 2018. textblob documentation. *Available at https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf* .

Maclean S. 2016. A new model for social work reflection: whatever the weather. *Professional Social Work* **1**:28–29.

Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, Pfetsch B, Heyer G, Reber U, Häussler T. 2018. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Communication Methods and Measures* **12(2–3)**:93–118 DOI 10.1080/19312458.2018.1430754.

Mathur A, Kubde P, Vaidya S. 2020. Emotional analysis using Twitter data during pandemic situation: COVID-19. In: *2020 5th international conference on communication and electronics systems (ICCES)*. Piscataway: IEEE, 845–848.

Mbwogge M. 2021. Mass testing with contact tracing compared to test and trace for the effective suppression of COVID-19 in the United Kingdom: systematic review. *JMIRx Med*. DOI 10.2196/27254.

McCormick TH, Lee H, Cesare N, Shojaie A, Spiro ES. 2017. Using Twitter for demographic and social science research: tools for data collection and processing. *Sociological Methods & Research* **46(3)**:390–421 DOI 10.1177/0049124115605339.

Medhat W, Hassan A, Korashy H. 2014. Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal* **5(4)**:1093–1113 DOI 10.1016/j.asej.2014.04.011.

Mohammad SM, Turney PD. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* **29(3)**:436–465 DOI 10.1111/j.1467-8640.2012.00460.x.

Mustaqim T, Mustaqim T, Umam K, Muslim MA, Muslim MA. 2020. Twitter text mining for sentiment analysis on government's response to forest fires with vader

lexicon polarity detection and k-nearest neighbor algorithm. *Journal of Physics: Conference Series* **1657(3)**:24–32 DOI 10.1088/1742-6596/1567/3/032024.

Nguyen D, Liakata M, DeDeo S, Eisenstein J, Mimno D, Tromble R, Winters J. 2020. How we do things with words: analyzing text as social and cultural data. *Frontiers in Artificial Intelligence* **3**:62 DOI 10.3389/frai.2020.00062.

NHS England. 2021. Data for contact tracing. London, UK: UK Health Security Agency.

Nikolenko SI, Koltcov S, Koltsova O. 2017. Topic modelling for qualitative studies. *Journal of Information Science* **43(1)**:88–102 DOI 10.1177/0165551515617393.

Ofqual. 2020. Awarding GCSE, AS & A levels in summer 2020: interim report. *Available at https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report*.

Pak A, Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. 1320–1326.

Park J, Ciampaglia GL, Ferrara E. 2016. Style in the age of instagram: predicting success within the fashion industry using social media. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work amp; social computing, CSCW '16*. New York, NY, USA: Association for Computing Machinery, 64–73 DOI 10.1145/2818048.2820065.

Pokharel BP. 2020. Twitter sentiment analysis during Covid-19 outbreak in Nepal. *SSRN Electronic Journal* **33(1)**:126–127 DOI 10.2139/ssrn.3624719.

Post G, Visser V, Buis J. 2017. 13. Reflection. In: *Academic skills for interdisciplinary studies*. Amsterdam, Netherlands: Amsterdam University Press, 116–123.

Prihatini PM, Prihatini P, Suryawan IK, Mandia I. 2018. Feature extraction for document text using Latent Dirichlet Allocation. *Journal of Physics: Conference Series* **953**:12–47 DOI 10.1088/1742-6596/953/1/012047.

Rehùøek R, Sojka P. 2011. Gensim–statistical semantics in Python. *Available at https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf* .

Ribeiro MT, Wu T, Guestrin C, Singh S. 2020. Beyond accuracy: behavioral testing of NLP models with checkList. ArXiv preprint. arXiv:2005.04118.

Rimmer A. 2021. *Sixty seconds on... the pingdemic*. London, UK: British Medical Journal Publishing Group.

Rish I. 2001. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3*, 41–46.

Roesslein J. 2009. Tweepy documentation. Online. 5 *Available at http://tweepy.readthedocs.io/en/v3*.

Rout JK, Choo K-KR, Dash AK, Bakshi S, Jena SK, Williams KL. 2018. A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research* **18(1)**:181–199 DOI 10.1007/s10660-017-9257-8.

Sailunaz K, Dhaliwal M, Rokne J, Alhajj R. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining* **8(1)**:1–26 DOI 10.1007/s13278-017-0479-5.

Samuel G, Roberts S, Fiske A, Lucivero F, McLennan S, Phillips A, Hayes S, Johnson S. 2021. COVID-19 contact tracing apps: UK public perceptions. *Critical Public Health* **32(1)**:1–13.

Saura JR, Ribeiro-Soriano D, Saldaña PZ. 2022. Exploring the challenges of remote work on Twitter users' sentiments: from digital technology development to a post-pandemic era. *Journal of Business Research* **142**:242–254 DOI 10.1016/j.jbusres.2021.12.052.

Schofield A, Mimno D. 2016. Comparing apples to apple: the effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics* **4**:287–300 DOI 10.1162/tacl_a_00099.

Sengers P, McCarthy J, Dourish P. 2006. Reflective HCI: articulating an agenda for critical practice. In: *CHI'06 extended abstracts on Human factors in computing systems*. 1683–1686.

Sengupta S. 2019. What are academic subreddits talking about? A comparative analysis of r/Academia and r/Gradschool. In: *Conference companion publication of the 2019 on computer supported cooperative work and social computing, CSCW '19*. New York, NY, USA: Association for Computing Machinery, 357–361 DOI 10.1145/3311957.3359491.

Sivalakshmi P, Kumar PU, Vasanth M, Srinath R, Yokesh M. 2021. COVID-19 vaccine public sentiment analysis using Python's Textblob approach. *International Journal of Current Research and Review* **13(11)**:166–172 DOI 10.31782/ijcrr.2021.sp218.

Smith H. 2020. Algorithmic bias: should students pay the price? *AI & Society* **35(4)**:1077–1078 DOI 10.1007/s00146-020-01054-3.

Song S, Min J, Kim H, Min K. 2019. Topic modeling to mind illegal compensation for occupational injuries. *European Journal of Public Health* **29(Supplement_4)**:ckz186–ckz317 DOI 10.1093/eurpub/ckz186.

Srinivasan B, Kumar D. 2019. Flock the similar users of Twitter by using latent dirichlet allocation. *International Journal of Scientific & Technology Research* **8(11)**:1421–1425.

Stine RA. 2019. Sentiment analysis. *Annual Review of Statistics and Its Application* **6**:287–308 DOI 10.1146/annurev-statistics-030718-105242.

Tang C-L, Liao J, Wang H-C, Sung C-Y, Cao Y-R, Lin W-C. 2020. Supporting online video learning with concept map-based recommendation of learning path. In: *Extended abstracts of the 2020 CHI conference on human factors in computing systems, Chi Ea '20*. New York, NY, USA: Association for Computing Machinery, 1–8 DOI 10.1145/3334480.3382943.

Teh Y, Jordan M, Beal M, Blei D. 2004. Sharing clusters among related groups: hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems* **17**:1385–1392.

Ulfa MA, Irmawati B, Husodo AY, yudo Husodo A. 2018. Twitter sentiment analysis using Na"ive bayes classifier with mutual information feature selection. *Journal of Computer Science and Informatics Engineering (J-Cosine)* **2(2)**:106–111 DOI 10.29303/jcosine.v2i2.120.

van Atteveldt W, van der Velden MA, Boukes M. 2021. The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures* **15(2)**:121–140 DOI 10.1080/19312458.2020.1869198.

Viera AJ, Garrett JM. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine* **37(5)**:360–363.

Villena-Román J, Garcıa-Morera J. 2013. TASS 2013 workshop on sentiment analysis at SEPLN 2013: an overview. In: *Proceedings of the TASS workshop at SEPLN*. 112–125.

Vyas V, Uma V. 2018. An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. *Procedia Computer Science* **125**:329–335 DOI 10.1016/j.procs.2017.12.044.

Wang Q, Saha K, Gregori E, Joyner D, Goel A. 2021. Towards mutual theory of mind in Human-AI interaction: how language reflects what students perceive about a virtual teaching assistant. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery.

Watanabe K. 2021. Latent semantic scaling: a semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures* **15(2)**:81–102 DOI 10.1080/19312458.2020.1832976.

Williams SN, Armitage CJ, Tampe T, Dienes K. 2021. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study. *Health Expectations* **24(2)**:377–385 DOI 10.1111/hex.13179.

Woodfield K, Morrell G, Metzler K, Blank G, Salmons J, Finnegan J, Lucraft M. 2013. Blurring the Boundaries? New social media, new social research: developing a network to explore the issues faced by researchers negotiating the new research landscape of online social media platforms. Southampton, UK: National Centre for Research Methods (NCRM) Networks for Methodological Innovation.

Yang S, Zhang H. 2018. Text mining of Twitter data using a latent dirichlet allocation topic model and sentiment analysis. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* **12(7)**:525–529.

PeerJ Computer Science

Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: a study on governmental preventive measures

Muhammad Faisal Ali, Rabia Irfan and Tahira Anwar Lashari

Department of Computing, School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan

ABSTRACT

Sentiments are the key factors that lead to influence our behavior. Sentiment analysis is a technique that analyzes people's behaviors, attitudes, and emotions toward a service, product, topic, or event. Since 2020, no country has remained untouched by COVID-19, and the governing bodies of most countries have been applying several anti-pandemic countermeasures to combat it. In this regard, it becomes tremendously important to analyze people's sentiments when tackling infectious diseases similar to COVID-19. The countermeasures taken by any country to control the pandemic leave a direct and crucial impact on each sector of public life, and every individual reacts to them differently. It is necessary to consider these reactions to implement appropriate messaging and decisive policies. Pakistan has done enough to control this virus's spread like every other country. This research aims to perform a sentimental analysis on the famous microblogging social platform, Twitter, to get insights into public sentiments and the attitudes displayed towards the precautionary steps taken by the Government of Pakistan in the years 2020 and 2021. These steps or countermeasures include the closure of educational institutes, suspension of flight operations, lockdown of business activities, enforcement of several standard operating procedures (SOPs), and the commencement of the vaccination program. We implemented four approaches for the analysis, including the Valence Aware Dictionary and sEntiment Reasoner (VADER), TextBlob, Flair, and Bidirectional Encoder Representations from Transformers (BERT). The first two techniques are lexicon-based. Flair is a pre-trained embedding-based approach, whereas BERT is a transformer-based model. BERT was fine-tuned and trained on a labeled dataset, achieving a validation accuracy of 92%. We observed that the polarity score kept varying from month to month in both years for all countermeasures. This score was analyzed with real-time events occurring in the country, which helped understand the public's sentiment and led to the possible formation of a notable conclusion. All implemented approaches showed independent performances. However, we noticed from the classification results of both TextBlob and the fine-tuned BERT model that neutral sentiment was dominant in the data, followed by positive sentiment.

Submitted 20 June 2022
Accepted 28 December 2022
Published 31 March 2023

Corresponding author
Muhammad Faisal Ali,
mali.mscs18seecs@seecs.edu.pk

Academic editor
Yu-Dong Zhang

Additional Information and Declarations can be found on page 29

DOI **10.7717/peerj-cs.1220**

Copyright
2023 Ali et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Data Mining and Machine Learning, Data Science, Network Science and Online Social Networks, Text Mining, Sentiment Analysis

Keywords COVID-19, Coronavirus, Sentiment analysis, Pakistan, Public health policies, Preventive measures, Corona SOPs

INTRODUCTION

The COVID-19 pandemic, initially caused by SARS-CoV-2, also known as the coronavirus pandemic, has been spreading worldwide. The virus was first identified in a seafood marketplace in Wuhan, China, in late 2019 (*Velavan & Meyer, 2020*) and has infected half a billion people around the globe. Despite taking various steps to contain it, the virus spread dramatically, and the WHO declared this outbreak a pandemic on March 11, 2020. As of April 11, 2022, almost 500 million people have been affected, and 6 million precious lives have been lost worldwide during this pandemic. We can obtain the absolute and final statistical insights once the pandemic is over (*Hamzah et al., 2020*); however, it is one of the deadliest outbreaks in history. The coronavirus pandemic is not the first pandemic, and the world has faced different outbreaks and pandemics in different periods, such as SARS, MERS-Cov, and Ebola. The community has dealt with all such situations, and their efforts have stood fruitful (*Elder et al., 2020*). COVID-19 has affected every aspect of life and led to various social and economic repercussions worldwide. Society faced supply shortages, including food, medicine, and other necessities of life. Many businesses faced the worst situations and got affected by the measures taken by the authorities to combat this deadly pandemic. The misinformation or partial information through media, like the promotion of false propaganda, led a few people to panic buying and escalated tension in various other aspects (*Islam et al., 2020*). In this instability, no one can ignore the contribution and role of technology. It has helped the world in different and effective ways in making crucial and medically necessary decisions to contain the virus (*Soliman & Tabak, 2020*). The historical data has allowed researchers to define new boundaries for combating such fatal diseases on a large scale (*Pan et al., 2020*). Mass media and social media have been proven as the most significant data sources and contributors, more than ever before, in analyzing major trends. Along with other various uses, intriguingly, such data is also utilized for the sentiment and perception analysis of the public over time.

Sentiment analysis is instrumental in contributing to finding human behaviors present on social media applications. Various real-life domains such as business, health, tourism, education, and many more have utilized such analyses. Different recommender systems are performing very efficiently based on the outcomes from sentiment analysis. It can help analyze real-time trends and choose the right path per those trends (*Chaudhary & Naaz, 2017*). A very large-scale population can be monitored and tracked at a meager cost using sentiment analysis (*Choi et al., 2017*). Hence, it is a potent tool to help us see the community's general perception. In short, the role and contribution of computer technologies cannot be ignored when tackling deadly diseases or pandemics. These technologies have provided numerous opportunities to fight such outbreaks (*Goldschmidt, 2020*), especially determining the public's perception of social and mass media. Multiple research studies have shown that many outbreaks could have been contained more effectively and quickly if social media data had been considered (*Singh, Singh & Bhatia, 2018*). The COVID-19 pandemic has been a controversial topic on social media. Based on the recent events during the pandemic, sentiment analysis can help better understand the public's sentiment about these events and implement a policy that can keep up with the

public's perception. Past events have shown that whenever a pandemic or similar emergency has occurred, it has profoundly affected people (*Vahia, Jeste & Reynolds, 2020*). Such situations can be controlled, but there are many reasons why this process of overcoming may be delayed or slowed down; one is panic buying (*Islam et al., 2021*). It becomes alarming when the community feels fear and doesn't follow the precautions the concerned authorities give because people are either ill-informed or misinformed about crucial information. It just adds fuel to the situation (*Erku et al., 2021*). In such cases, it's imperative to study the perception of the general people from time to time so that the proper guidelines can be conveyed where needed.

Recently, when COVID-19 emerged in Wuhan, China, countries worldwide started to gear up to contain this deadly virus. Pakistan is a country that shares its border with China, so different preventive measures were introduced and implemented from time to time to combat the pandemic in the country. Pakistan has so far faced four dangerous waves of this virus. As of April 17, 2022, more than 1.5 million people have been affected, and over 30 thousand have died. The lives of ordinary citizens have been affected by the tough decisions to deal with these situations. The importance of considering public sentiment is discussed in detail; the motivation and objective behind this study are to analyze public sentiments for each of the countermeasures taken by the Government of Pakistan in containing the spread of the coronavirus. Some significant contributions of the underlying study are as follows:

- Five comprehensive datasets about different preventive measures taken to combat COVID-19 are collected.
- Four approaches are implemented on each dataset to analyze the sentiment about all preventive measures.
- The sentiment during four COVID-19 waves at different times in the country is analyzed and compared.
- The results obtained from the analysis can help concerned authorities in the future in conveying the proper guidelines and implementing decision policies following the sentimental tone of the public.

The article is organized as follows: "Related Work" presents a detailed overview of work related to the study under discussion. "Materials and Methods" provides the comprehensive characteristics of the dataset and the models' frameworks used during the research. The obtained results are presented in "Results," whereas "Discussion" contains a brief discussion about the critical findings of the underlying study. In the end, "Conclusion" concludes the presented case with some future directions.

RELATED WORK

The research works under discussion below contain studies from various geographical zones and different implemented techniques, analyzing the data collected through multiple sources and for different periods.

Liu & Liu (2021) studied the public behavior toward the COVID-19 vaccine on Twitter data posted from November 1, 2020, to January 31, 2021. First, they used the VADER (*Hutto & Gilbert, 2014*) to determine the sentiment from the data and LDA (*Blei, Ng & Jordan, 2003*) for the topic modeling from positive and negative Tweets. They extracted five main topics for the positive sentiment (trial results, administration, life, information, and efficacy) and five for the negative category (trial results, conspiracy, trust, effectiveness, and administration). They also performed temporal analysis using the PELT algorithm (*Killick, Fearnhead & Eckley, 2012*). At last, they performed geographic analysis using the Twitter Geolocation package. This study was conducted when the vaccine was not available to the community. In another study, *Zhang et al. (2021)* extracted tweets between February 24, 2020, and October 14, 2020. They analyzed four different American cities and the same number of Canadian cities. This study determined how people are reacting to the coronavirus overall in different timespans and identified the public sentiment about three primary anti-epidemic measures "the lockdown," "mask," and "vaccine." They compared the results from the cities of both countries to each other. NCR (*Mohammad & Turney, 2013*) was used to classify each Tweet's emotions, and VADER to categorize the sentiments. They concluded that the number of infections with the coronavirus and the public sentiment was correlated.

Surano, Porfiri & Rizzo (2021) analyzed a Twitter dataset consisting of 1.3 million tweets from different regions of the USA from January 21 to May 31, 2020. This analysis was done on the containment measure "lockdown" and the change in perception and sentiment about it. They considered "education" and "wealth indicators" and collected the relevant data from US Census Bureau. For sentiment classification, this study used VADER and classified tweets into positive, negative, and neutral categories. Furthermore, two different time series were created for each region (state) and the DC area, with a daily fraction of positive and negative tweets. Multiple time series were collected and divided into three sections: before the occurrence of the pandemic, from its occurrence to the first peak, and from its rise to the end of May 2020. They studied each region's positive and negative portions and calculated the averages. In research on Singapore's "Circuit Breaker" *Ridhwan & Hargreaves (2021)* did a sentiment analysis on Twitter data from February 1, 2020, to August 31, 2020. They used VADER for the sentiment analysis, and tweets' compound score was considered for the classification. A pre-trained RNN was used to determine the emotions from the tweets. For topic modeling, LDA and GSDMM were used. They tried to find a correlation between the changes in these sentiments and the occurrence of real-life events in the country. Another study by *Dubey (2020)* performed sentiment analysis on the Twitter data extracted from 12 different countries from March 11, 2020, to March 31, 2020. This research used the NRC model, which classifies the text based on eight emotions defined by *Plutchik (1980)*, to categorize the sentiments of the tweets in the form of expressed emotion rather than the numerical value like in VADER. This study had two analysis phases; in the first phase, they discovered that Belgium was the city where 63% of the tweets showed positive sentiment and only 37% were negative. China was the country containing the most negative tweets from its population. In the second

phase, they performed the analysis with the emotions and observed that the USA, France, and China contained the highest no. of tweets with an "anger" tone.

Samuel et al. (2020) performed sentiment analysis on data posted on Twitter from February 2020 to March 2020, targeting the keyword "corona." They tried identifying the public sentiments related to the outbreak using COVID-related tweets and R statistical software with sentiment analysis packages. This study was limited to the United States, where they showed insights into fear sentiment over time in the country as the pandemic was approaching its peak level. They performed a textual analysis aided by the visualization of descriptive text. Moreover, they used two machine learning classification models, the naïve Bayes classifier and logistic regression, and compared their performance on textual analysis of tweets of different lengths. The naïve Bayes method performed well on short tweets with an accuracy of 91%, while logistic regression gave a reasonable classification accuracy of 74%. Both models performed relatively weakly for tweets containing longer text. In another study conducted in the Philippines for sentiment analysis about the COVID-19 vaccine, (*Villavicencio et al., 2021*) extracted tweets weekly from 1st to 31st March 2021. These extracted tweets were 11,974 in total and were both in English and Tagalog language. This was the first month of the vaccination drive in the country. This data was manually annotated with three different sentiments: positive, negative, and neutral. They used the TF-IDF approach for vectorization and the naïve Bayes classifier for classification. They evaluated the model with k-fold cross-validation with k = 10, which achieved an accuracy of 81.77%. (*Machuca, Gallardo & Toasa, 2021*) performed a sentiment analysis on Twitter data posted from January to July 2020. They extracted the data for every month with the keyword "#coronavirus" and studied using TF-IDF vectorizer, which assigns a weightage to a word within the text. They used a pre-trained binary logistic regression to classify the tweets. The applied model obtained an overall accuracy of 78.57%. They discovered that 46% of the tweets showed negative sentiments, and 54% were positive ones.

Garcia & Berton (2021) performed a topic detection and sentiment analysis on English and Portuguese tweets. They got the top ten topics under discussion and analyzed the data discussed on Twitter over time. To avoid the weak performance on short text with a reason for the lack of word co-occurrence information, they used GSDMM for topic modeling. At the same time, they combined three recent embedding models for feature extraction, *i.e.*, sBERT (*Reimers & Gurevych, 2019*), mUSE (*Yang et al., 2019*), and FastText (*Bojanowski et al., 2017*). For English tweets, they used CrystalFeel (*Gupta & Yang, 2018*) for sentiment analysis. They also compared performance in both languages in this study. *Singh, Jakhar & Pandey (2021)* performed a sentimental analysis on two different datasets, one containing tweets from India and the other containing data from worldwide. Their dataset contained tweets over 80 days, from January 20 to April 25, 2020. They used mRMR (*Radovic et al., 2017*) for feature selection and the BERT model (*Devlin et al., 2018*) to classify tweets. They achieved 93.89% validation accuracy and concluded that the people of India had a positive sentiment in those days compared to the rest of the world. In another study, *Kumar (2022)* performed a sentiment analysis on the public sentiment in India over time in multiple phases of lockdown and its gradual lifting. This study used the Sentiment140 dataset for

the training and a manually annotated tweets dataset for testing purposes. A hybrid deep learning approach that consisted of BiLSTM and CNN was used for the sentiment classification. The primary purpose of adding CNN along with the BiLSTM was feature extraction. This hybrid model showed an overall accuracy of 90%.

Zhou et al. (2021) conducted sentiment analysis research in Beijing, China. They collected Sina Weibo data from the initial days of the virus, from January 13 to February 28, 2020, and at the time of its re-emergence, from June 4 to July 20, 2020. They divided the data into subsets based on three countermeasures, lockdown, test-trace-isolate strategy, and the ban on the gatherings. The LDA was used for the topic extraction and Baidu for the categorizing sentiments. They observed that the people of China greatly supported the test-trace-isolate measure (60%) in the initial days of the pandemic. This support increased to the highest level (90%) during the re-emergence of Corona cases in the country. In a study conducted by *Tan et al. (2021)* on Sina Weibo, several public posts from different accounts were collected and analyzed across 31 Chinese provinces from January 2018 to December 2020. They followed the Zipf distribution to normalize the number of Weibo posts per account and used a professional Chinese sentiment API named Tencent NLP Product to determine the sentiment. In major, they focused on the economic indicators and analyzed the public sentiment before and during the pandemic. They discovered that the pandemic had had severe effects on public mental health.

Above, we have discussed multiple studies from various geographical regions analyzing data collected from different timeframes. It can be observed that there is no comprehensive research conducted in the region of Pakistan. Pakistan is a very close neighbor to China, where this deadly virus emerged, and has taken multiple steps to combat the pandemic. Although a few surveys have been done in Pakistan to check the public sentiment regarding COVID-19, all such surveys were done on a tiny scale where a range of 500–1,000 people participated. As the sentiment analysis and an overview of the public perception from social media, where the public has a significant presence, is essential to be extracted, which can help in implementing various needy steps according to the situation in the future, this study is performed with the mentioned purpose. An overview with salient points of studies discussed above is presented in Table 1.

MATERIALS AND METHODS

Various studies conducted by different authors have adopted different steps in performing sentiment analysis; however, the central workflow remains common, which is depicted more thoroughly in Fig. 1.

First, a source platform is selected on which one wants to perform the analysis. For example, when it comes to social media analysis, a user can choose from various sources like Twitter, Instagram, Facebook, Reddit, *etc.* The selection of a platform depends upon the purpose, and it varies from analysis to analysis. After finalizing a platform, a user needs to extract the data. Every platform has its different ways of data gathering. Users can target one or more specific keywords to extract the information they want. The extracted data will also be in various forms, such as posts, tweets, news, or text. Third, comes data pre-processing, which includes multiple steps such as cleaning; the removal of stopwords,

Table 1 An overview of studies discussed in related work.

Sr. #	Year	Data Source	Timespan	Region	Analysis approach
1	2021	Sina Weibo	Jan 18–Dec 20	Mainland China	Tencent API
2	2021	Sina Weibo	Jan 20–Feb 20	China	Baidu (OSS)
3	2021	Twitter	Jan 20–Apr 20	India *vs* RoTW	Automated
4	2022	Twitter	Mar 20–onward	India	Automated
5	2020	Twitter	Feb 20–Mar 20	USA	Automated
6	2021	Twitter	Nov 20–Jan 21	USA (State & National Level)	Rule-based
7	2021	Twitter	Apr 20–Aug 20	USA & Brazil	Automated
8	2021	Twitter	Feb 20–Oct 20	USA & Canada	Rule-based
9	2021	Twitter	Jan 20–May 20	USA	Rule-based
10	2021	Twitter	Feb 20–Aug 20	Singapore	Rule-based
11	2021	Twitter	Mar 2021	Philippines	Automated
12	2020	Twitter	Jan 20–Jul 20	Multiple countries	Automated
13	2020	Twitter	Mar 2020	Multiple countries	Rule-based

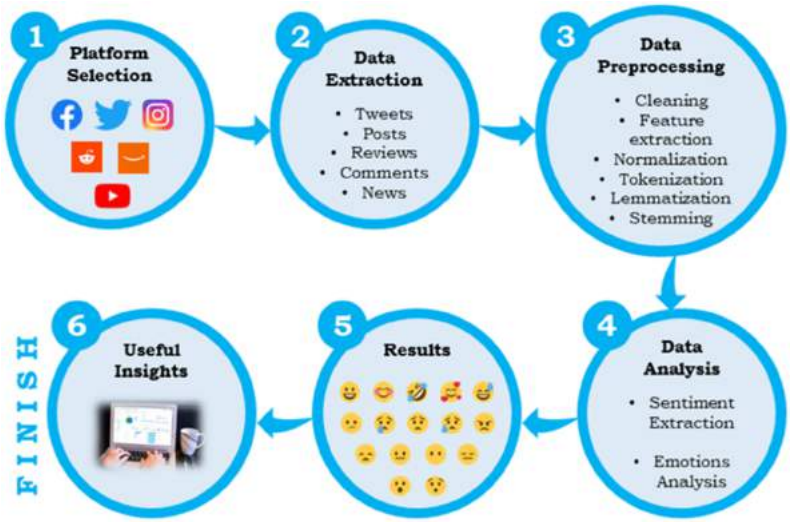


Figure 1 Workflow of sentiment analysis or opinion mining.
Full-size DOI: 10.7717/peerj-cs.1220/fig-1

duplicates, and unwanted portions of text; feature extraction; structuring of the data and extraction of meaningful portions, normalization; converting all data to the same form such as lowercase, tokenization; converting the text to some keywords or tokens which contain some specific meanings, stemming; removal of last few characters of a word, and lemmatization; converting the text form to a meaningful base form. The pre-processed data is passed for sentiment analysis in the final phase, and some valuable findings are extracted.

Datasets collection

The data was extracted from the social platform Twitter using the Snscrape library (github.com/JustAnotherArchivist/snscrape). Twitter has the edge over other social media platforms when analyzing a specific topic or a trend. It's the platform that networks ideas rather than people like Facebook. A post on Facebook may contain different topics because there is no such restriction on the text length, but on the other hand, Twitter allows only 280 characters per tweet. So, a tweet is usually relevant to a specific topic. Snscrape is an instrumental library that enables one to extract tweets without any personal API keys. It features extensive search options that permit customizable searches and can return thousands of records in minutes. Twitter has complex phenomena about fetching tweets by location. There are currently two ways to do so; either if a tweet is geotagged or if a site is mentioned in the user's profile. As per Twitter, a significantly less portion of tweets is geotagged, so it is not an optimal way. On the other side, many people have a location in their bio, but they can fill it up with whatever they want. Some are very kind and write properly, such as "Karachi, Pakistan," but a few write things like "My Dad's House." It was found that some Twitter algorithms work in a way in its advanced search where they can find tweets by location if it is a part of a user's bio and assume that all tweets coming from this profile will be tweeted from or around that specific location, or if the tweet is geotagged. There are two ways to make such an algorithm work for us; we can either provide coordinates of a location with a radius, and it will return with tweets posted in that radius, or we can use the "near" feature, such as "near: Karachi." As per our requirement, we needed the data from a country, *i.e.*, Pakistan. It was not an excellent way to provide coordinates of a country and a radius to collect the data, as it can result in inaccurate records. So, the best approach was to scrape tweets by cities and yield the best possible results. The data was extracted from the 100 most populated cities of Pakistan, presented in Table 2, targeting a radius of 50 km for each city.

Pakistan has adopted various preventive measures from time to time in fighting COVID-19 (*Jabeen et al., 2020*), including the closure of educational institutes, suspension of flight operations, lockdown of business activities, enforcement of several standard operating procedures (SOPs), and the commencement of the vaccination program. These were the steps that affected public life the most. Data were extracted for each of those countermeasures with the help of relevant keywords in specific periods, whose details are given in Table 3. These datasets are available for the community to use for further research (*Ali, 2022*).

Data pre-processing

The process of data pre-processing is fundamental in data analysis. Here unwanted pieces of data are removed, and only a useful portion is kept. There can be several unwanted things in a tweet, like hashtags, mentioned user handles, *etc.* Table 4 represents the steps adopted while pre-processing the extracted raw text. It is pertinent to mention that Twitter hashtags can also contain some sentiments sometimes. In our case, most of them were related to COVID-19, such as #Coronavirus, #COVID, #Lockdown, #Vaccination, *etc.*, and are present in almost every tweet. Such hashtags had no specific encoded sentiment

Table 2 Most populated cities of Pakistan from which data was extracted.

Sr.#	City	Sr.#	City	Sr.#	City
1	Abbotabad	34	Islamabad	67	Mirpur Mathelo
2	Ahmedpur East	35	Jacobabad	68	Multan
3	Arif Wala	36	Jaranwala	69	Muridke
4	Attock	37	Jatoi	70	Muzaffarabad
5	Badin	38	Jhang	71	Muzaffargarh
6	Bahawalnagar	39	Jhelum	72	Narowal
7	Bahawalpur	40	Kabal	73	Nawabshah
8	Bhakkar	41	Kamalia	74	Nowshera
9	Bhalwal	42	Kamber Ali Khan	75	Okara
10	Burewala	43	Kāmoke	76	Pakpattan
11	Chakwal	44	Kandhkot	77	Peshawar
12	Chaman	45	Karachi	78	Quetta
13	Charsadda	46	Kasur	79	Rahim Yar Khan
14	Chiniot	47	Khairpur	80	Rawalpindi
15	Chishtian	48	Khanewal	81	Sadiqabad
16	Dadu	49	Khanpur	82	Sahiwal
17	Daharki	50	Khushab	83	Sambrial
18	Daska	51	Khuzdar	84	Samundri
19	Dera Ghazi Khan	52	Kohat	85	Sargodha
20	Dera Ismail Khan	53	Kot Abdul Malik	86	Shahdadkot
21	Faisalabad	54	Kot Addu	87	Sheikhupura
22	Ferozwala	55	Kotri	88	Shikarpur
23	Ghotki	56	Lahore	89	Sialkot
24	Gojra	57	Larkana	90	Sukkur
25	Gujranwala	58	Layyah	91	Swabi
26	Gujranwala Cantt.	59	Lodhran	92	Tando Adam
27	Gujrat	60	Mandi Bahauddin	93	Tando Allahyar
28	Gwadar	61	Mansehra	94	Tando Mohd. Khan
29	Hafizabad	62	Mardan	95	Taxila
30	Haroonabad	63	Mianwali	96	Turbat
31	Hasilpur	64	Mingora	97	Umerkot
32	Hub	65	Mirpur	98	Vehari
33	Hyderabad	66	Mirpur Khas	99	Wah Cantonment
100	Wazirabad				

which could add something to the overall sentiment of a tweet, so we removed them during pre-processing.

Approaches selected for the analysis

There were four approaches implemented during the underlying study. The first two, VADER and TextBlob, were rule-based techniques. Third, a pre-trained embedding-based

Table 3 An overview of datasets extracted against each countermeasure.

Countermeasures	No. of tweets extracted	Time period (Months)	Starting from (Month)	Extracted until (Month)
Closure of educational institutes	672,465	19	Mar 2020	Sep 2021
Suspension of flight operations	159,418	22	Mar 2020	Dec 2021
Lockdown of business activities	646,679	22	Mar 2020	Dec 2021
Standard operating procedures (SOPs)	214,019	22	Mar 2020	Dec 2021
Vaccination program	162,548	15	Dec 2020	Feb 2022

Table 4 Pre-processing steps taken during the analysis.

Sr. #	Pre-processing step	Description	Example
1	Hashtags	Hashtags help extract data relevant to a specific topic but are not so useful after getting data scraped. So, these hashtags were removed from the data.	#COVID-19, #Vaccination, #Lockdown
2	User handles	On Twitter, each user has a unique username, such usernames help specify a person but have no use in the text analysis. So, all user handles were removed.	@OfficialNcoc, @fslsltn, @WHO
3	URLs	A tweet may contain an external link or an attachment in the form of a hyperlink. All such things provide no benefit in text analysis and were removed.	bit.ly/3mcQZEs
4	Useless characters	A few useless characters were also removed from the text, just to make it neater and cleaner.	@, #, |, \n, \t, \r
5	Multiple spaces	Multiple spaces occurred when the above pieces of texts were removed from the data, these were also removed to get data in a normal form.	More than one space
6	Null values	Null or empty records were also removed.	Any empty record
7	Duplicates	All duplicates amongst the records were discarded, so that each tweet should be unique and present a meaningful insight.	Any duplicate record
8	Punctuations	Some techniques do not require punctuations as they work with the tokens or lexicons only, so punctuations were removed while implementing such techniques.	?, !, ;, :, -, …
9	Stopwords	Stopwords do not add much meaning in lexicon-based approaches. These were removed while applying such approaches.	My, we, it, to, from, be, has, have, do, so

model, "Flair" (*Akbik, Blythe & Vollgraf, 2018*), was used. In the end, a fine-tuned transformer-based model, "BERT," was considered for sentiment classification. This fine-tuned BERT was first trained on annotated data and then used on the unseen data extracted from Twitter discussed above.

VADER

VADER (github.com/cjhutto/vaderSentiment) is a rule-based sentiment analysis approach specifically sensitive to presumptions expressed on social media. It uses pre-annotated lexicons that are labeled as positive, negative, or neutral and categorize any sentence based on its lexicons' sentiments. VADER not only tells us about polarity but also provides information on how much positivity, negativity, and neutrality there is in a sentence. VADER has some limitations, but this approach is still considered a good option for

sentiment analysis from social media. Before applying VADER, data was pre-processed, and the first seven steps in Table 4 were adopted. The punctuations and stopwords were not removed from the data because VADER does consider such elements in the text. It is case-sensitive and gives different sentiment polarity for "english" and "English." VADER provides a probability distribution of sentiment for a sentence. It tells us the probability of a positive, negative, or neutral sentence. The sum of this distribution is always one. In addition, it provides an overall polarity score known as the "compound" score, which was selected for deciding on a tweet's sentiment in this study. A threshold was set for the compound score such that if the score was greater or equal to 0.05, the tweet was labeled as positive; if the score was less than or equal to −0.05, it was labeled as negative; otherwise, it was considered as a neutral one. VADER was applied to all datasets mentioned above, and the results are discussed in detail.

TextBlob

TextBlob (github.com/sloria/TextBlob) is a simple Python library that works on a lexicon-based strategy. Here, the individual sentiments of lexicons constructing a sentence determine the sentence's sentiment. It uses "nltk" and allows us to perform extensive textual analysis and operations, including sentiment analysis. While implementing this approach returns two factors: polarity and subjectivity. The polarity score lies between −1 and 1, where −1 shows the negativity in the sentiment of a sentence and 1 implies the positivity. The negation is used to reverse the polarity of a sentence. The amount of "personal opinion" and actual data in a text is measured by "subjectivity." If a sentence has a higher value of subjectivity, it focuses more on "personal opinion" rather than the facts in the text. TextBlob returns zero if the sentence has no such word with a polarity score in the "nltk" lexicons-dataset because it calculates a sentiment score based on the weighted average of all the words present in a sentence. For the most part, TextBlob performs well in simple situations. Before applying TextBlob, data was pre-processed. Unlike VADER, the punctuations and stopwords were also removed from the data because TextBlob does not consider these elements. A threshold was also set here for the polarity score such that if the score was greater than zero, the tweet was labeled as a positive tweet; if the score was less than zero, the same was labeled as negative, and it was considered as a neutral one if the polarity score stood zero. This approach was also implemented on all datasets, and the results are further discussed in detail.

Flair embeddings

Flair (github.com/flairNLP/flair) is an open-sourced simple library whose framework is built on top of PyTorch, which itself is a state-of-the-art deep learning framework. The flair library comes with plenty of useful features. It contains valuable word embeddings like ELMo (*Joshi, Peters & Hopkins, 2018*), GloVe (*Pennington, Socher & Manning, 2014*), character embedding, *etc*. Models like Flair are usually known as embedding-based models. The "Flair Embeddings" are the central embedding provided by this library which works based on contextual string embeddings. These use the inner states of different pre-trained models and provide new word embeddings. Using internal rules of multiple character

models can ensure consideration of the word's context. Like the other two previous approaches, data were pre-processed, and the first seven steps mentioned in Table 4 were adopted. The punctuations and stopwords were kept in the text to ensure that the model understands the context of tweets. This model returned two things: the sentiment label of the sentence and the model's confidence in labeling the sentence. As the flair model is trained on IMDb reviews, it could return only two sentiments, *i.e.*, positive or negative. The third sentiment, "neutral," could not be analyzed using flair embedding.

Bidirectional encoder representations from transformers (BERT)

NLP has been achieving marvelous success in developing entities that can understand how best to represent the data, especially the text data, and understand the hidden meanings and relationships between different portions of the data. In addition, the NLP community has contributed several practical components that can be freely utilized in models developed for various tasks in the domain of NLP. The release of the BERT model is one of the most recent accomplishments amongst such achievements. This achievement is the start of a new era for NLP. BERT is a stack full of pre-trained Transformer encoders. This model has two different sizes: BERT Base and BERT Large. The base version has 12 encoder layers, while the large version has 24 layers. The model also has large feed-forwarded networks (hidden units), 768 in the base version and 1,024 in the large version, and attention heads, 12 and 16 for the base and large versions, respectively. As input, a sequence of words is given to BERT. Each encoder layer uses self-attention, then sends the results through a feed-forward network before passing them on to the next encoder layer. The first token is provided with a special token, CLS, which stands for classification. Another special token, SEP, which stands for separator, is used and added to separate the different sentences in the text. Each position generates a vector whose size is known as the "hidden_size." This size is 786 in the base version and 1,024 in the large version. The generated vector is then passed as an input to any classifier we want. The BERT model is fine-tuned and used to classify the sentiment in the study under discussion. The details of the process of fine-tuning are as follows:

Training dataset

The dataset for training the BERT model is open-sourced and publicly available on GitHub (*Jadeja, 2022*). It contains 200,000 tweets related to COVID-19 collected in 2 years, 2020 and 2021. The dataset is passed through a few steps of pre-processing. The hashtags, user handles, URLs, duplicates, and null records are removed from the data. It is a highly balanced dataset with an equal number of records from all classes. There are three sentiment classes: positive, negative, and neutral; amongst them, 66,700 tweets are positive, 66,660 are negative, and 66,640 are neutral. There is a total of seven columns in the data: timestamp of the tweet, tweet ID, username, the content of the tweet, cleaned content of the tweet, sentiment label, and sentiment score. In the sentiment score, "2" is assigned to the positive label, "1" to the neutral, and "0" to the negative sentiment. The dataset was split into a 70–30% ratio for training and testing. The training set contains 140,000 tweets (70%), and the testing set contains 60,000 (30%). Both sets are also balanced and have an

equal number of each class. In the training set, 46,690 tweets are positive, 46,662 tweets are negative, and 46,648 records are neutral, while in the testing set, there are 20,010 positive, 19,998 negatives, and 19,992 neutral tweets. The balanced data ensures that a model would not lead to misleading results and give equal importance to each class present in the dataset.

Model architecture

First, the class "AutoTokenizer" and "TFBertModel" from Transformers were used to load the tokenizer and the BERT model from the Hugging Face library. In this study, a "bert-base-cased" (huggingface.co/bert-base-cased) tokenizer and model were used to process our data to shape it into a form required by the BERT. The "bert-base-cased" is a base version of the BERT model and is case-sensitive. In our case, the tokens' length was just around 150, so to be on the safer side, the maximum length for the tokenizer was set to 160. The special tokens were added, and the "truncation," "padding," and "attention_mask" were set to True. As a result, the "input_ids" and "attention_mask" of the modeled data were returned. While building our model, we provided two input layers for the BERT layer. After that, the embeddings have hidden states. Then, using "GlobalMaxPooling1D" to build CNN layers, a dense layer of 128 units with "Relu" as the activation function, a dropout layer to avoid overfitting, a second dense layer with 32 units, and again "Relu" as the activation function, and at last, a dense layer with "3" units because there were three classes in our data with a "Sigmoid" function which led to the output of our function.

Model training

For the compilation of the model built, Adam optimizer was used. The "learning_rate" was set to 5e-5 with a decay of 0.01, which is a bit slower but works perfectly. Since we were analyzing the multiclass problem, the loss was calculated from the "CategoricalCrossEntropy," and the balanced accuracy was taken from "CategoricalAccuracy" to include the average accuracy of all the classes in the data. First, the model was trained on ten epochs with a batch size of 16. The model was generalized well on data after ten epochs, but we trained on ten more epochs with the same batch size to achieve even better performance. Fig. 2 represents the training history of twenty epochs.

Model evaluation

The model showed a validation accuracy of 89.02% after the first 10 epochs, which increased to almost 92% after the further ten epochs. The validation loss calculated during these phases was 0.3093 and 0.2315, respectively. After 20 epochs, the model had been generalized ideally on the data. The trained model was applied to all datasets discussed in the same section, and classification was performed. The results obtained are discussed from all angles in detail in the next section.

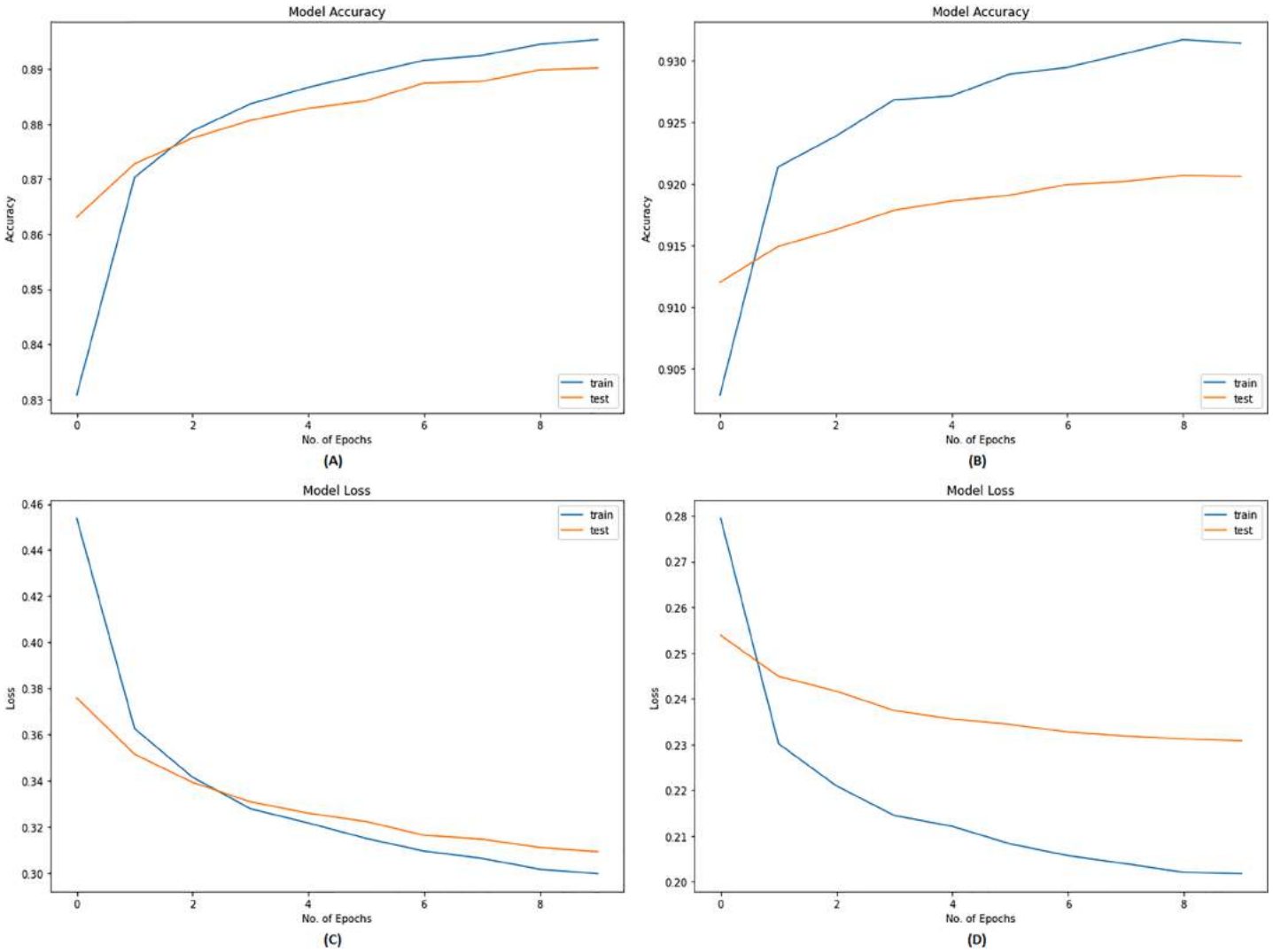


Figure 2 Training history of BERT model. (A) Accuracy graph after training on first 10 epochs. (B) Accuracy graph after training on next 10 epochs. (C) Loss graph after training on first 10 epochs. (D) Loss graph after training on next 10 epochs.

Full-size DOI: 10.7717/peerj-cs.1220/fig-2

RESULTS

Each approach showed its independent classification results on each dataset, whose illustrations and valuable insights are presented in this section. Table 5 represents the sentiments' distributions classified by all four models.

Closure of educational institutes

Figure 3 depicts the sentiment percentages distribution per month about the closure of educational institutes. It can be observed from the classification results of VADER that the positive sentiment has a leading presence overall in the data than both the negative and neutral sentiments. In the year 2020, December, June, and April are the months with the highest percentage of positive, negative, and neutral sentiments, respectively, whereas, in

Table 5 Sentiment percentage distribution done by all models.

Preventive measures	Model implemented	Sentiment percentages distribution		
		Positive	Negative	Neutral
Closure of educational institutes	VADER	43.66%	24.19%	32.15%
	TextBlob	39.36%	16.26%	44.37%
	Flair	48.23%	51.77%	N/A
	BERT	33.55%	18.98%	47.47%
Suspension of flight operations	VADER	41.30%	31.26%	27.44%
	TextBlob	38.93%	18.13%	42.93%
	Flair	38.48%	61.52%	N/A
	BERT	34.41%	28.43%	37.16%
Lockdown and closure of all business activities	VADER	45.36%	24.79%	29.85%
	TextBlob	38.93%	18.13%	42.93%
	Flair	43.85%	56.15%	N/A
	BERT	36.36%	20.57%	43.07%
Enforcement of standard operating procedures	VADER	46.18%	27.95%	25.87%
	TextBlob	46.58%	16.71%	36.71%
	Flair	34.88%	65.12%	N/A
	BERT	39.02%	24.28%	36.70%
Commencement of vaccination program	VADER	44.31%	22.42%	33.27%
	TextBlob	39.14%	16.13%	44.72%
	Flair	32.33%	67.67%	N/A
	BERT	39.54%	19.36%	41.10%

the year 2021, the months of February, May, and June contained the highest percentage of the same sentiments. Amongst both years, the positive sentiment was highest in February 2021. May 2021 was the month with the highest negative sentiment, while the neutral percentage was at its highest in April 2020. In the data classified by TextBlob, it can be noticed that the neutral sentiment has a leading presence overall in the data than positive and negative. In the year 2020, December, November, and March are the months with the highest percentage of neutral, positive, and negative sentiment, respectively, whereas, in the year 2021, the months of January, March, and May contain the highest percentage of the same sentiments. In both years, the neutral sentiment was highest in December 2020. November 2020 was the month with the highest positive sentiment, while the negative percentage was at its highest in May 2021.

The Flair model is trained on the IMDB reviews, and since there was no neutral sentiment, only two classes were considered here. We can see that the negative sentiment has a higher presence in the data than the positive. In 2020, October and March were the months with the highest positive and negative sentiment percentages, respectively. In the year 2021, the months of February and April contain the highest percentage of the same sentiments. In both years, positive sentiment was highest in February 2021, while the negative percentage was at its highest in March 2020. It can be observed from the

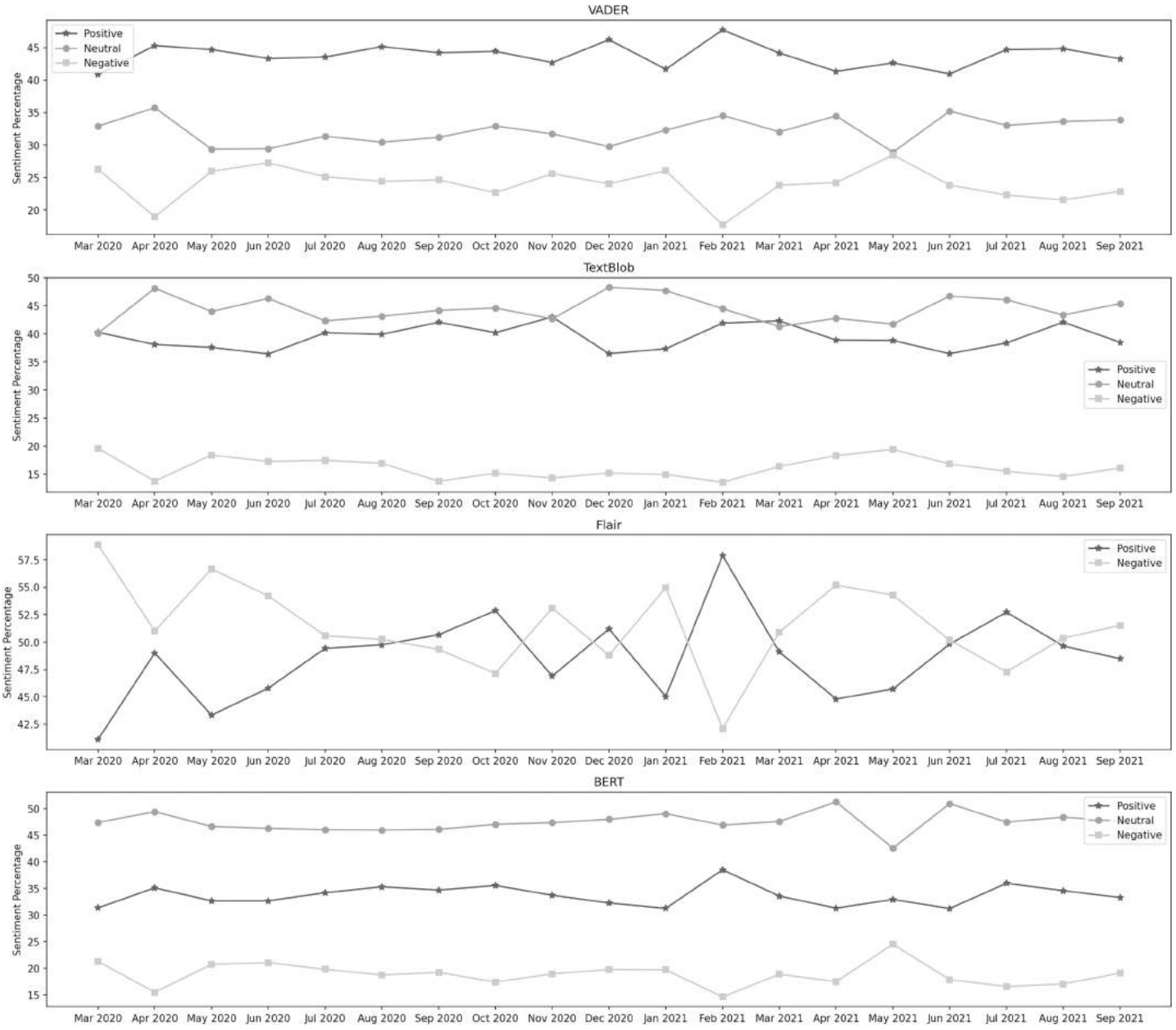


Figure 3 **Sentiment percentage per month about the closure of educational institutes.** Full-size DOI: 10.7717/peerj-cs.1220/fig-3

classification results of BERT that the neutral sentiment has a leading presence overall in the data than positive and negative. In the year 2020, April, October, and March are the months with the highest percentage of neutral, positive, and negative sentiment, respectively, whereas, in the year 2021, the months of April, February, and May contain the highest percentage of the same sentiments. During both years, the neutral sentiment was highest in April 2021. February 2021 was the month with the highest positive sentiment, while the negative percentage was at its highest in May 2021.

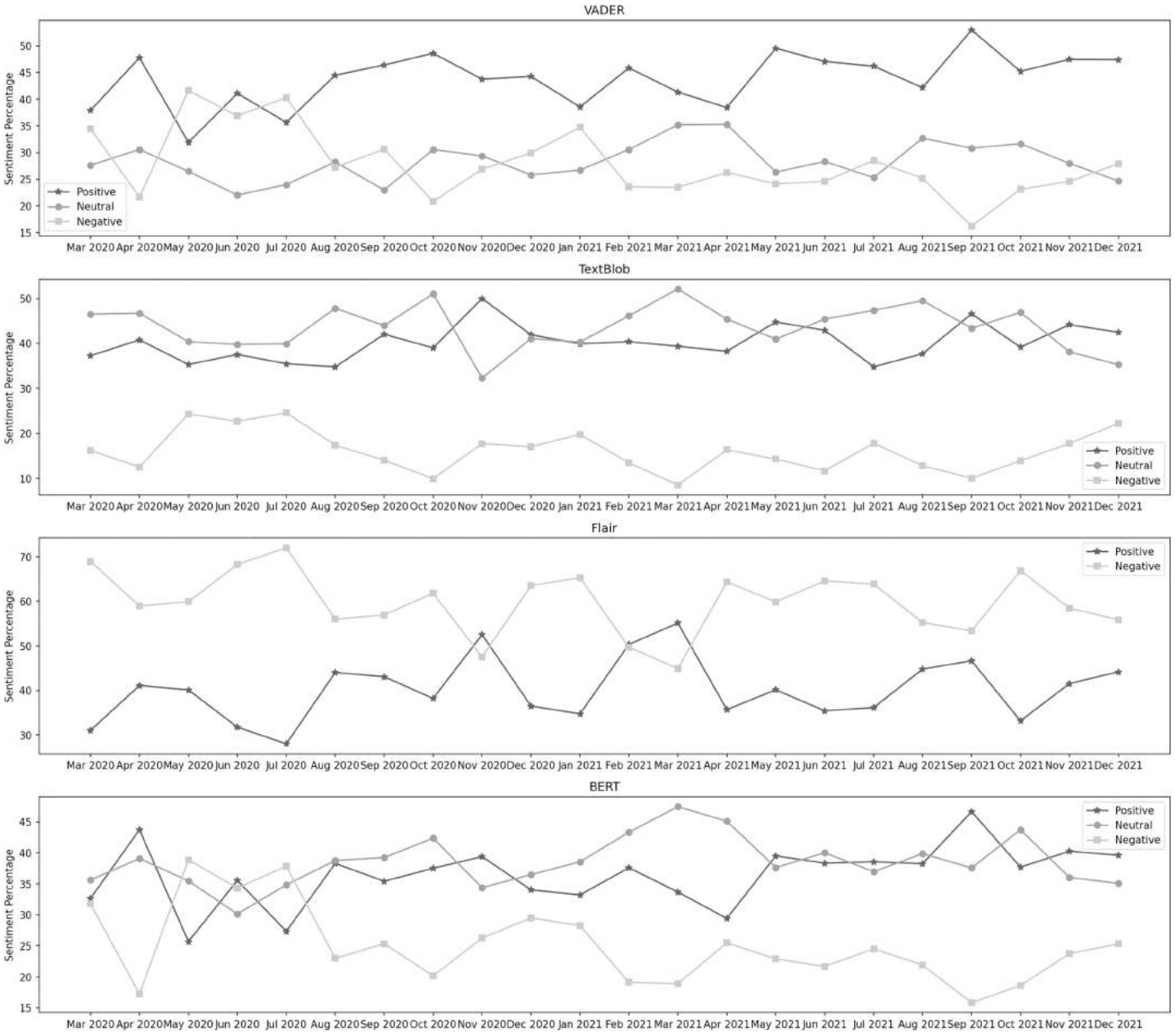


Figure 4 **Sentiment percentage per month about the suspension of flight operations.** Full-size DOI: 10.7717/peerj-cs.1220/fig-4

Suspension of flight operations

Figure 4 depicts the sentiment percentages distribution per month about the suspension of flight operations. It can be observed from the classification results of VADER that the positive sentiment has a leading presence overall in the data than negative and neutral. In the year 2020, October, May, and April are the months with the highest percentage of positive, negative, and neutral sentiments, respectively, whereas, in the year 2021, the months of September, January, and April contain the highest percentage of the same

sentiments. Among both years, positive sentiment was highest in September 2021. April 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in May 2020. In the data classified by TextBlob, the neutral sentiment has a higher overall presence than positive and negative. In the year 2020, October, November, and July are the months with the highest percentage of neutral, positive, and negative sentiment, respectively, whereas, in the year 2021, the months of March, September, and December contain the highest percentage of the same sentiments. During both years, the neutral sentiment was highest in March 2021. November 2020 was the month with the highest positive sentiment, while the negative percentage was at its highest in July 2021.

The results from the Flair model show that the negative sentiment has a higher presence in the data than the positive. In 2020, November and July were the months with the highest percentage of positive and negative sentiments, respectively. In 2021, March and October contained the highest percentage of the same sentiments. In both years, positive sentiment was highest in March 2021, while the negative percentage was at its highest in Jul 2020. It can be noticed from the classification results of BERT that the neutral sentiment has a leading presence overall in the data than positive and negative. In the year 2020, October, April, and May are the months with the highest percentage of neutral, positive, and negative sentiment, respectively, whereas, in the year 2021, the months of March, September, and January contain the highest percentage of the same sentiments. During both years, the neutral sentiment was highest in March 2021. September 2021 was the month with the highest positive sentiment, while the negative percentage was at its highest in May 2020.

Lockdown and closure of all business activities

Figure 5 depicts the sentiment percentages distribution per month about the lockdown and closure of business activities. In the data classified by VADER, it can be observed that the positive sentiment has a leading presence overall in the data than negative and neutral. In the year 2020, December, June, and October are the months with the highest percentage of positive, negative, and neutral sentiments, respectively, whereas, in the year 2021, the months of November, May, and October contain the highest percentage of the same sentiments. Among both years, positive sentiment was highest in November 2021. October 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in June 2020. The results from TextBlob show that neutral sentiment has a higher presence in the data than positive and negative. In the year 2020, March, December, and June are the months with the highest percentage of neutral, positive, and negative sentiments, respectively, whereas, in the year 2021, the months of October, November, and April contain the highest percentage of the same sentiments. During both years, the neutral sentiment was highest in October 2021. November 2021 was the month with the highest positive sentiment, while the negative percentage was at its highest in June 2020.

From the classification results of Flair, we can observe that the negative sentiment has a more leading presence overall in the data than the positive. In 2020, December and March were the months with the highest positive and negative sentiment percentages, respectively. In 2021, November and April contained the highest percentage of the same

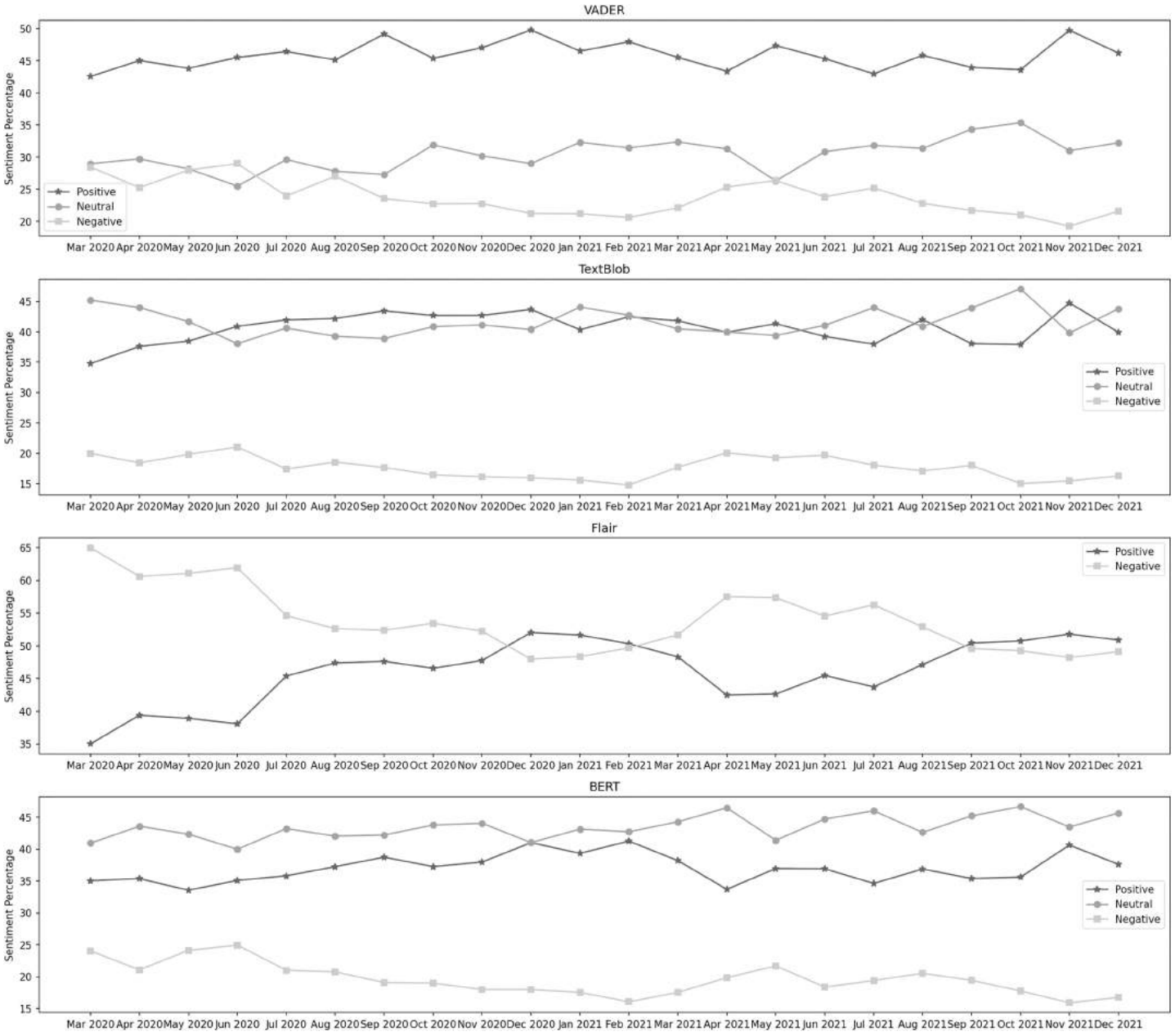


Figure 5 **Sentiment percentage per month about the lockdown and closure of business activities.**

Full-size DOI: 10.7717/peerj-cs.1220/fig-5

sentiments. In both years, positive sentiment was highest in December 2020, while negative percentage was highest in March 2020. In the data classified by BERT, it can be noticed that the neutral sentiment has a leading presence overall in the data than positive and negative. In 2020, November, December, and June were the months with the highest neutral, positive, and negative sentiment percentages. In 2021, October, February, and May contained the highest percentage of the same sentiments. During both years, the neutral

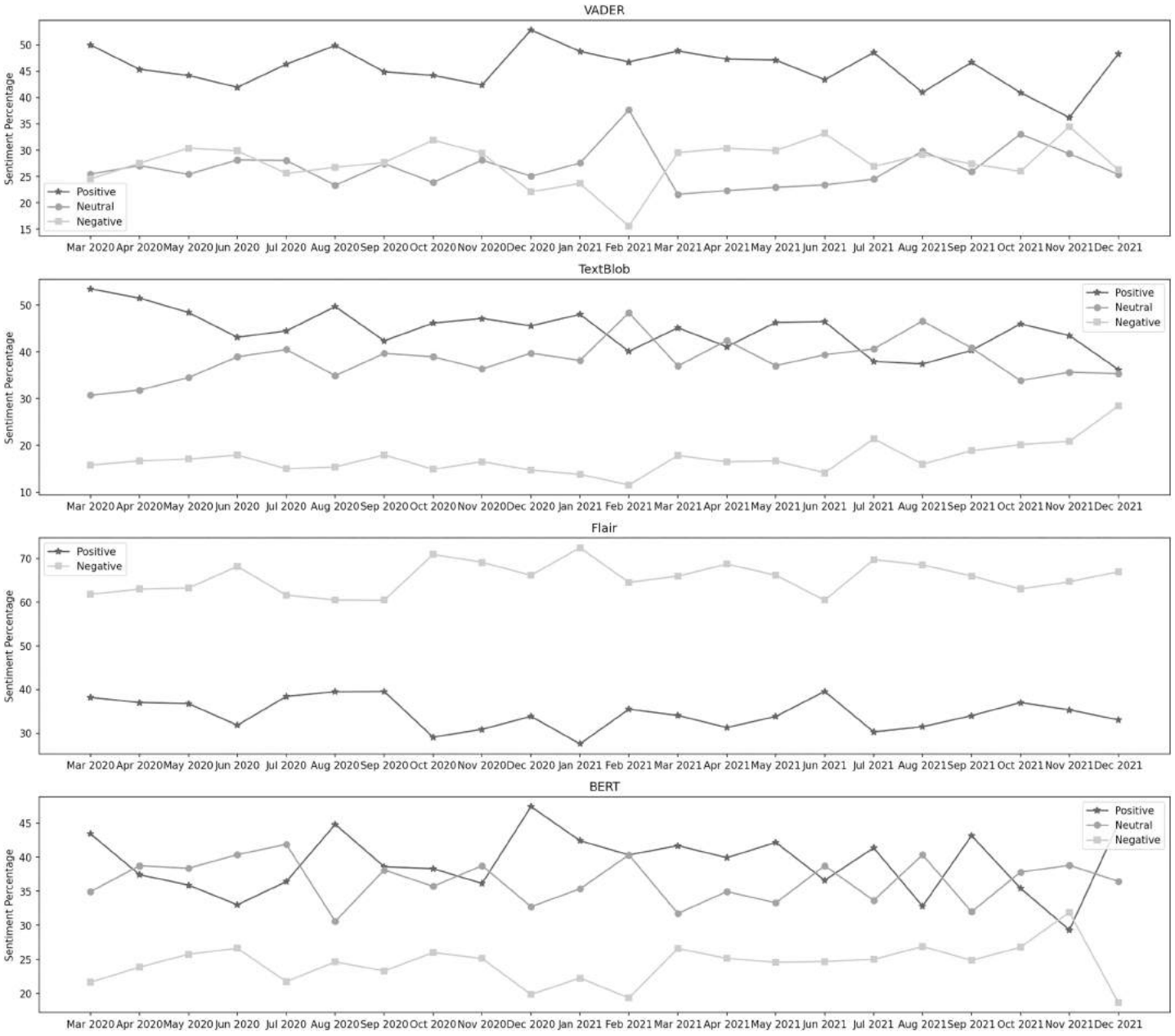


Figure 6 **Sentiment percentage per month about the enforcement of standard operating procedures.**
Full-size DOI: 10.7717/peerj-cs.1220/fig-6

sentiment was highest in October 2021. February 2021 was the month with the highest positive sentiment, while the negative percentage was at its highest in June 2020.

Standard operating procedures (SOPs)

Figure 6 depicts the sentiment percentages distribution per month regarding the enforcement of standard operating procedures. From VADER's results, we can see that the positive sentiment has a leading presence overall in the data than negative and neutral. In

the year 2020, December, June, and October are the months with the highest percentage of positive, negative, and neutral sentiments, respectively, whereas, in the year 2021, the months of March, February, and November contain the highest percentage of the same sentiments. Among both years, positive sentiment was highest in December 2020. February 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in November 2021. The distribution by TextBlob shows that the positive sentiment has a leading presence overall in the data than neutral and negative. In the year 2020, March, September, and June are the months with the highest percentage of positive, negative, and neutral sentiments, respectively, whereas, in the year 2021, the months of January, December, and February contain the highest percentage of the same sentiments. Among both years, positive sentiment was highest in March 2020. February 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in December 2021.

It can be observed from the classification results of Flair that the negative sentiment has a leading presence overall in the data than the positive. In 2020, September and October were the months with the highest positive and negative sentiment percentages, respectively. In 2021, June and January contained the highest percentage of the same sentiments. In both years, positive sentiment was highest in June 2021, while the negative percentage was at its highest in January 2021. BERT's results show that the positive sentiment has a higher presence in the data than neutral and negative. In the year 2020, July, December, and June are the months with the highest percentage of neutral, positive, and negative sentiment, respectively, whereas, in the year 2021, the months of August, December, and November contain the highest percentage of the same sentiments. During both years, the neutral sentiment was highest in July 2020. December 2020 was the month with the highest positive sentiment, while the negative percentage was at its highest in November 2021.

Vaccination program

Figure 7 depicts the sentiment percentages distribution per month about the vaccination program. In the data classified by VADER, the positive sentiment has a higher overall presence than negative and neutral. In the first seven months, from December 2020 to June 2021, December, January, and February are the months with the highest percentage of positive, negative, and neutral sentiment, respectively, whereas, in the next 8 months, from July 2021 to February 2022, the months of January, November, and December contain the highest percentage of the same sentiments. Amongst all these months, positive sentiment was highest in January 2022. December 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in January 2021. It can be observed from the classification results of TextBlob that the neutral sentiment has a leading presence overall in the data than positive and negative. In the first 7 months, from December 2020 to June 2021, January, March, and December are the months with the highest percentage of positive, negative, and neutral sentiment, respectively, whereas, in the next 8 months, from July 2021 to February 2022, the months of January, September, and December contain the highest percentage of the same sentiments. Amongst all these

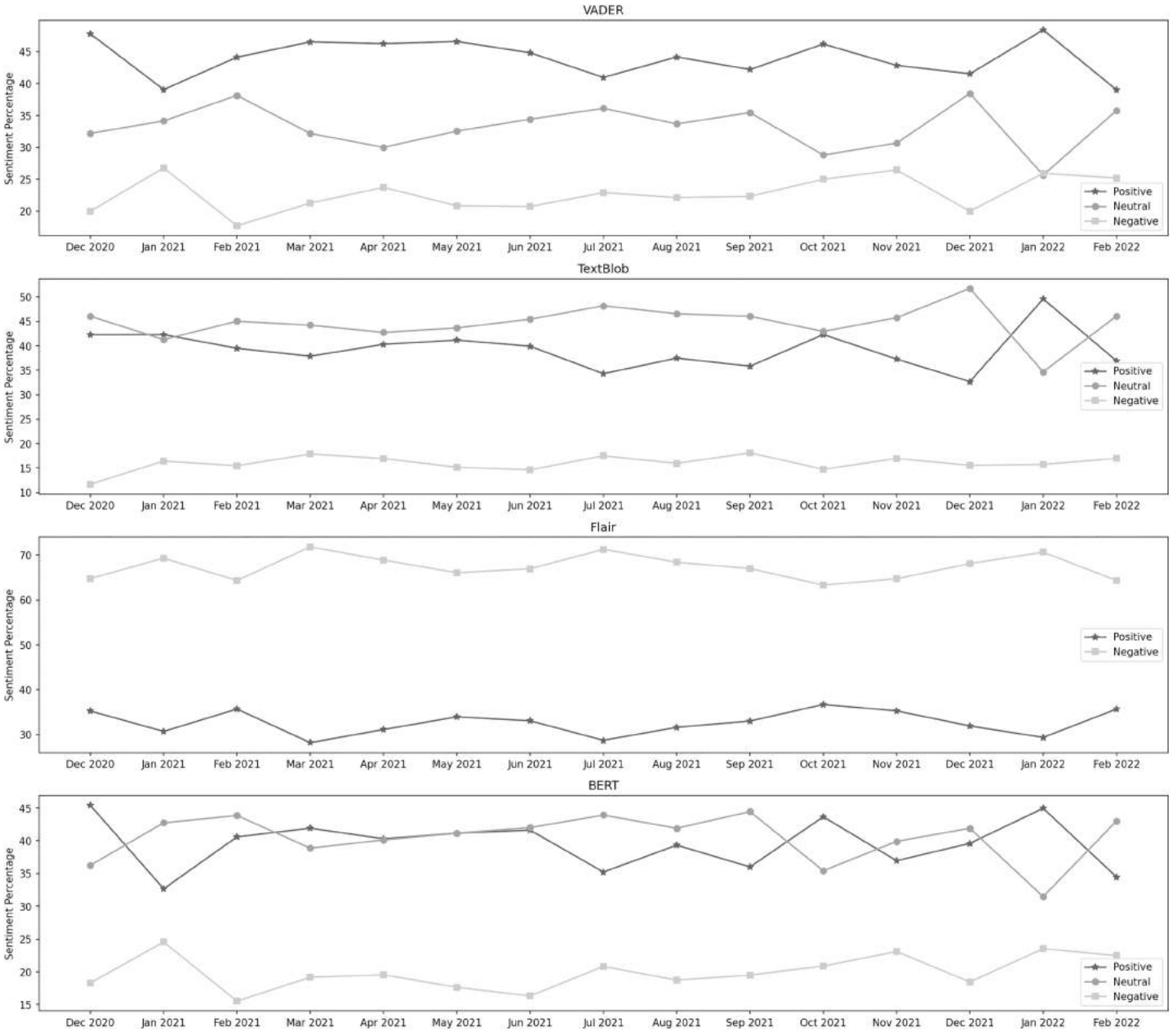


Figure 7 **Sentiment percentage per month about the vaccination program.** Full-size DOI: 10.7717/peerj-cs.1220/fig-7

months, positive sentiment was highest in January 2022. December 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in September 2021.

The results from the Flair model show that the negative sentiment has a leading presence overall in the data than positive. In the first 7 months, from December 2020 to June 2021, February and March are the months with the highest percentage of positive and negative sentiment, respectively, whereas, in the next 8 months, from July 2021 to February

2020, the months of October, and July contain the highest percentage of the same sentiments. Amongst all these months, positive sentiment was highest in October 2021, while the negative percentage was at its highest in March 2021. We can notice from BERT's results that the neutral sentiment has a leading presence overall in the data than positive and negative. In the first 7 months, from December 2020 to June 2021, December, January, and February are the months with the highest percentage of positive, negative, and neutral sentiment, respectively, whereas, in the next 8 months, from July 2021 to February 2022, the months of January, January, and September contain the highest percentage of the same sentiments. Among all these months, positive sentiment was highest in December 2020. September 2021 was the month with the highest neutral sentiment, while the negative percentage was at its highest in January 2021.

Summary

The implemented approaches are based on different techniques, which have been discussed earlier in detail. Each of them has its capabilities and limitations. VADER is considered a powerful tool for social media analysis; the lexicon vocabulary used by VADER is not updated regularly, and it is challenging for this approach to evaluate the slang of this modern era, which is used very extensively in social media and change day by day. TextBlob is also a rule-based technique that calculates the polarity score of individual terms and determines the overall sentiment of the sentence. Flair embeddings are trained on IMDb reviews. The sentimental tone of a review and a social media post is usually very different, and such pre-trained embeddings cannot be accurately used for the sentimental analysis of a social media platform toward a specific topic. Another crucial aspect to focus on in sentiment analysis is the context of a sentence. The context must be preserved while determining the sentimental attitude of the text. Considering all these factors, building a custom model and fine-tuning it for a specific purpose becomes essential. A fine-tuned model can overcome such drawbacks and give the most accurate results. In the study under discussion, a fine-tuned BERT model is used, which is discussed in the previous section in detail. As this model is trained on a labeled Twitter dataset and contains all the capabilities of a state-of-the-art transformers-based model, the results shown by this custom-trained model were considered the most accurate and taken to a noticeable conclusion. These results were analyzed and compared at the time of four COVID-19 waves in the country during the years 2020 and 2021.

By the end of 2021, Pakistan faced four dangerous coronavirus waves. The first wave was noticed from March 2020 till the end of July 2020, the second wave remained from the start of November 2020 till the end of January 2021, the third wave lasted from March 2021 till the end of May 2021, and the fourth and last wave in these 2 years was observed from July 2021 to September 2021. The preventive measures under discussion were implemented across the country in different phases during these four waves. The time-to-time public sentiment about these measures is discussed as follows:

1. All educational institutes, including schools, colleges, universities, and madrasas, were closed across the country. The examinations got postponed for the time being. Later,

most institutes, excluding primary and middle schools, were allowed to resume their activities online. After a period, fifty percent of strength got permitted to attend face-to-face classes. These developments continued to occur in various situations depending on the pandemic scenario in the country (*The Express Tribune, 2020b*; *The International News, 2020*; *Ary News, 2020b*). The overall sentiment remained neutral in the data about the closure of educational institutes. During the first wave, in March 2020, almost 48% of the sentiment distribution was neutral, while the positive and negative distribution was 31% and 21%, respectively. In April 2020, the negative sentiment decreased to 16%, while the positive and neutral percentages increased to 35% and 49%, respectively. In May 2020, there was an increase in the negative sentiment, whereas both other percentages got slightly decreased. This distribution level remained almost the same for the next 2 months. At the start of the second wave, in November 2020, the positive, neutral, and negative sentiment percentage was 34%, 47%, and 19%, respectively. The positive sentiment decreased in the next 2 months, while neutral and negative sentiments increased. At the end of the second wave, the percentage of the same sentiments was 31%, 49%, and 20%, respectively. During the third wave, in March 2021, 48% of the sentiment was neutral, while 33% and 19% were positive and negative, respectively. In the next month, a significant increase was observed in the neutral sentiment. In the third month, there was a massive decrease in the neutral percentage, whereas the negative sentiment increased. At the time of the last wave, in July 2021, the neutral sentiment percentage was 47% which went up to one point in August 2021 and again came down to 47% in the month of September 2021. The positive sentiment was noticed at 36% in July 2021. It gradually decreased in the next 2 months and was found to be 33% in the last month of the fourth wave. The negative sentiment followed the opposite way and increased from 17% to 19% from the first to the last month, September 2021. Overall, public behavior remained neutral or positive about the closure of educational institutes throughout the period.

2. Following the other measures, all international and domestic flight operations got suspended. However, the officials arranged a few special flights to bring some nationals stuck overseas. These protocols kept changing following the policies set by the countries on the other side of the routes (*The Guardian, 2020*; *Kaleej Times, 2020*; *Business Today, 2020*). The public sentiment about the suspension of flight operations kept changing on a significant scale throughout the period. During the first wave, in March 2020, it was as follows: 35% neutral, 33% positive, and 32% negative. In April 2020, there was a massive hike in positive sentiment, which increased to 44%, and the neutral sentiment also increased to 39%. A substantial decrease was observed in the negative sentiment. In May 2020, the positive and neutral sentiments decreased to 26% and 35%, respectively, whereas the negative sentiment increased to 39%. In June 2020, the neutral distribution decreased to 30%. The negative distribution also got down to 34%, while the positive sentiment increased to 36%. In July 2020, the neutral and negative sentiments increased to 35% and 38%, respectively, and the positive percentage went down. In the second wave, in November 2020, the positive, negative, and neutral sentiment distributions

were 39%, 26%, and 35%, respectively. In December 2020, the same distribution went to 34%, 30%, and 36%. In the last month of the second wave, the neutral distribution slightly increased, whereas both other sentiments observed a slight decrease. During the third wave, in March 2021, the positive sentiment percentage was 34%, while the neutral and negative percentages were 47% and 19%, respectively. Next month, the negative percentage increased to 26%. In May 2021, a significant increase was noticed in positive sentiment. At the time of the last wave, in July 2021, the positive sentiment was 39% of the overall distribution, which went to 38% in August 2021 and increased to 47% in September 2021. The neutral distribution remained at 37%, 40%, and 37% in July, August, and September 2021, respectively. The neutral sentiment was 24% in July 2021 and gradually decreased in the next 2 months by the end of the fourth wave.

3. When the number of cases was getting higher daily, the federal and provincial governments imposed a strict lockdown at different levels. All public areas were closed, excluding pharmacies, grocery shops, and vegetable stores. The offices, wedding halls, parks, hotels, restaurants, and eateries were closed. The govt. also banned interprovince and intercity transport at different times. The country later opted for a smart-lockdown strategy; only hot areas were cordoned off wherever required (*The Express Tribune, 2020a*; *Arain, 2020*; *Ary News, 2020a*; *OyeYeah, 2020*; *Statesman, 2020*). It can be observed that the overall sentiment about the lockdown of all business activities remained neutral in both years. During the first wave, in March 2020, almost 41% of the sentiment distribution was neutral, while the positive and negative distribution was 35% and 24%, respectively. In April 2020, the negative sentiment decreased to 21%, while the neutral percentage increased to 44%. In May 2020, there was a decrease in positive and neutral sentiments. At the start of the second wave, in November 2020, the percentage of positive, neutral, and negative sentiment was 38%, 44%, and 18%, respectively. Next month, the positive percentage increased to 41%, while the neutral percentage decreased. In January 2021, a slight decrease in positive sentiment was observed, whereas the neutral sentiment increased slightly. During the third wave, in March 2021, 44% of the sentiment was neutral, while 38% and 18% were positive and negative, respectively. In April 2021, a significant decrease was observed in the positive sentiment. In May 2021, there was an increase in the positive and negative percentages, which went to 37% and 22%, respectively. At the time of the last wave, in July 2021, the neutral sentiment percentage was 46% which went down to 43% in August 2021 and again increased to 45% in September 2021. The positive sentiment was noticed at 35% in July 2021. It increased in the next month to 37% and was found to be 35% in the last month of the fourth wave. The negative sentiment slightly went up from 19% to 20% from July 2021 to August 2021 and remained the same in September 2021. The public sentiment remained neutral or positive about the lockdown of all business activities.
4. People were asked to comply with multiple SOPs set by WHO, including mask-wearing, hand sanitization, social distancing, *etc.* Several restrictions were also imposed against the non-compliant people (*Al-Jazeera, 2020*; *Dawn, 2020*; *Ahmed, 2020*; *Geo News, 2020a*; *Associated Press of Pakistan, 2020*). The public sentiment about the SOPs

remained positive or neutral throughout the period. During the first wave in March 2020, the sentiment distribution was 35% neutral, 43% positive, and 22% negative. A decrease in positive sentiment went to 37% in April. During May 2020, there was a slight decrease in positive and neutral sentiments. In June 2020, the positive distribution decreased to 35%. The neutral and negative percentages increased to 40% and 27%, respectively. In the last month of the first wave, the positive and neutral sentiments increased to 36% and 42%, respectively, and the negative percentage went significantly down to 22%. During the second wave, in November 2020, the positive, negative, and neutral sentiment distributions were 36%, 39%, and 25%, respectively. In December 2020, the same distribution went to 47%, 33%, and 20%. In the last month of the second wave, the positive distribution decreased, whereas both other sentiments observed an increase. During the third wave, in March 2021, the positive sentiment percentage was 42%, while the neutral and negative percentages were 32% and 26%, respectively. Next month, the neutral percentage increased to 35%, while both other sentiments' distributions decreased slightly. In May 2021, an increase was noticed in positive sentiment. The neutral and negative distributions reduced a bit. At the time of the fourth and last wave, in July 2021, the positive sentiment was 41% of the overall distribution, which went to 33% in August 2021 and increased to 43% in September 2021. The neutral distribution remained at 34%, 40%, and 32% in July, August, and September 2021, respectively. The neutral sentiment was 25% in July 2021; it increased to 27% in the next month, August 2021, and went again to 25% in September 2021 by the end of the fourth wave.

5. The initiation of the mass vaccination program was a significant relief for the public across the country. First of all, medical and paramedical staff was vaccinated. It was made available for the people, categorizing them with their age groups, and mandatory for every citizen above ten later (*Geo News, 2020b*; *Geo News, 2021*; *Dawn, 2021*; *Al-Jazeera, 2021*; *The Express Tribune, 2021*). December 2020 and January 2021 were analyzed to test the sentiment about the COVID vaccine before it was publicly available. It can be observed that the sentiment remained positive and neutral throughout the period. In December 2020, the positive sentiment percentage was 46%, while the neutral and negative sentiment distributions were 36% and 18%, respectively. The positive sentiment decreased to 33% in the next month, while the neutral sentiment went up to 43%. Pakistan started its COVID-19 vaccination drive on February 02, 2021. It was first made available for the front-line medical and paramedical staff and later for the public. In February 2021, the sentiment distribution was as follows: positive 41%, neutral 44%, and negative 15%. The vaccine was made available to people over 60 on March 10, 2021. In March 2021, the positive sentiment slightly increased to 42%. People over 50 were allowed to get vaccinated from March 30, 2021. In April 2021, the percentage distribution was as follows: positive 40%, neutral 40%, and negative 20%. From May 03, 2021, the age limit was 40 for vaccination eligibility. In May 2021, the positive and neutral sentiments slightly increased to 41% each. The Government of Pakistan allowed people over 30 to get their vaccination shots from May 30, 2021, and after a few days,

from June 03, 2021, all over 18 years of age were allowed to get vaccinated. In June 2021, the sentiments were as follows: positive 42%, neutral 42%, and negative 16%. Next month, the positive sentiment got decreased to 35%. In August, the positive sentiment again increased and went to 39%. The neutral and negative percentages got slightly decreased. August 30, and 31, 2021, were the days when the highest number of vaccination doses were administrated to the public. Over the next 6 months, the positive and neutral sentiments remained dominant over the distribution.

DISCUSSION

All the implemented approaches were performed independently, and the results obtained from the state-of-the-art BERT model were considered the most accurate because of its capabilities to overcome the limitations of the other three approaches. Five preventive measures were selected for the analysis, including the closure of educational institutes, suspension of domestic and international flight operations, lockdown of all business activities, enforcement of standard operating procedures, and the immunization drive started by the Government of Pakistan. The public sentiment during the four waves of the coronavirus was analyzed, and it came under observation that the sentiments were neutral throughout the period about the closure of educational institutes. The individual distribution kept varying from month to month, but overall, it remained neutral, followed by an influx of positive sentiments. During the first wave, people's sentiments were inconsistent about the suspension of flight operations. Initially, it was positive, but the negative distribution spiked the next month. However, it kept switching between positive and neutral sentiments during the subsequent waves. The sentiments about the lockdown and closure of all business activities were noticed as neutral. December 2020 was the month during which the positive and neutral distributions were almost equal; otherwise, the percentage remained neutral. It was observed that the people also remained positive or neutral about enforcing standard operating procedures (SOPs) too. Only in November 2021 the negative distribution was above the positive one; however, this month's neutral sentiment was on top. Pakistani people showed the same sentiment about the vaccination program also. It remained positive or neutral during the 15 months considered for the analysis. Pakistan had never faced an emergency on such a large scale before the emergence of COVID-19. The countermeasures deployed by the governing authorities were nothing less than setting a new course of daily life for every individual. The measures analyzed in this study are the ones that have affected public life the most. The closure of educational institutes for such a long period and then shifting the majority of them to an online mode of action was not routine for the general public. The suspension of flight operations was also affecting the whole community. Many overseas Pakistanis and people working abroad got stuck in their respective countries, and there was no way to bring them back to Pakistan. Imposing the lockdown, which was shifted to a smart-lockdown strategy later, and fixing the business hours impacted the citizens extensively. Many SOPs, including wearing face masks and gloves, hands-sanitization, maintaining social distance, and others, were the things the public was facing for the first time. When COVID vaccines were in the

laboratory phase and not available to the community, there was a mixed reaction in public about them. Later, even during the vaccination campaign at the masses level, they faced vaccination hesitancy in some groups of people. As mentioned earlier, these were the steps the community met for the first time in life. The people were dying, and no one could predict anything about the pandemic. It can be expected public sentiments to vary at different levels during such a scenario. Based on the observations, we can say that the public showed appreciative, either positive or neutral, behavior towards the preventive measures taken by the Government of Pakistan over time containing the spread of the deadly coronavirus.

CONCLUSION

The research under discussion is performed to conduct a comprehensive analysis to observe the public sentiment about the countermeasures taken by the Government of Pakistan in combating the COVID-19 disease. Such analyses are considered especially important because the response from the public is a crucial factor when implementing state-level policies. Data from the microblogging social platform Twitter was analyzed during the years 2020 and 2021. Five comprehensive datasets about different preventive measures are collected, which include the closure of educational institutes, suspension of flight operations, lockdown of business activities, enforcement of several standard operating procedures (SOPs), and the commencement of the vaccination program. Four approaches, VADER, TextBlob, Flair, and BERT, are implemented for the analysis. The first two, VADER and TextBlob, are rule-based techniques, while Flair and BERT are automated or learning-based approaches. BERT is fine-tuned on pre-labeled data, which showed a validation accuracy of 92% after 20 epochs. The public sentiment at the time of four COVID-19 waves during the years 2020 and 2021 is compared, and it is noticed from the results of the state-of-the-art transformer-based BERT model that the people of Pakistan had either neutral or positive sentiments about the counter strategies implemented against the spread of the coronavirus. Although this study is a comprehensive one and answers many of the sought questions, it still has some limitations which can be considered for improvements while performing some related work in the future:

- This study is performed on English-language tweets, while most Pakistani people speak Urdu, the country's national language. A similar sort of analysis can be done on the data in the Urdu language, which can present the picture in a bigger frame.
- Some important factors such as emojis, emoticons, and emphasized words, like tooooo good, contribute to determining the degree of the sentiment of a text, especially on social media. They help us know about the sentiment level of a particular piece of text. A study can be proposed in the future, probing the effects of including and excluding such factors.
- The sentiment may have several reasons to be positive, neutral, or negative. To get much better insights, topic modeling, a process of determining the semantic structures from text, and aspect-based sentiment analysis, which is a text analysis method that divides

opinions into aspects and determines the sentiment associated with each, can be performed on all datasets discussed in this study.

- Sarcasm is an aspect that can affect sentiment analysis in various ways. A generic sentiment analysis study cannot consider sarcasm as well. A study can be carried out on detecting sarcasm in the discussed data.
- An aspect-based sarcasm detection can be a comprehensive and contributing angle in this scenario.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Muhammad Faisal Ali conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Rabia Irfan conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Tahira Anwar Lashari analyzed the data, authored or reviewed drafts of the article, also provided invaluable assistance with the reviews and comments given by the editor and reviewers, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data is available at Zenodo: mohdfaisalali. (2022). mohdfaisalali/public-sentiments-about-pakistan-s-anti-covid-measures: v.01 (v.01). Zenodo. https://doi.org/10.5281/zenodo.7360872.

REFERENCES

Ahmed A. 2020. Pakistani manufactured coronavirus testing kits, ventilators to hit markets soon. *Available at https://www.brecorder.com/news/584835/*.

Akbik A, Blythe D, Vollgraf R. 2018. Contextual string embeddings for sequence labeling. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 1638–1649.

Al-Jazeera. 2020. Pakistani doctors decry lack of supplies as lockdown looms. *Available at https://www.aljazeera.com/news/2020/3/23/pakistani-doctors-decry-lack-of-supplies-as-lockdown-looms*.

Al-Jazeera. 2021. Vaccine hesitancy in Pakistan heightens risk of COVID resurgence. *Available at https://www.aljazeera.com/news/2021/3/5/in-pakistan-vaccine-hesitancy-heightens-risk-of-covid-19-resurge*.

Ali MF. 2022. Twitter dataset of Pakistan's Anti-COVID measures. *Available at https://www.kaggle.com/datasets/muhammadfaisalali/twitter-dataset-of-pakistans-anticovid-measures*.

Arain AL. 2020. Pakistan: Karachi's province set for COVID-19 lockdown. *Available at https://www.aa.com.tr/en/asia-pacific/pakistan-karachis-province-set-for-covid-19-lockdown/1774971*.

Ary News. 2020a. CM Punjab usman buzdar announces lockdown in Punjab. *Available at https://arynews.tv/cm-punjab-usman-buzdar-lockdown-punjab/*.

Ary News. 2020b. KP govt announces to close all educational institutions amid coronavirus fears. *Available at https://arynews.tv/coronavirus-fears-kp-educational-institutions-closed/*.

Associated Press of Pakistan. 2020. NCOC to re-close public service sectors if SOPs compliance goes neglected. *Available at https://www.app.com.pk/national/ncoc-to-re-close-public-service-sectors-if-sops-compliance-goes-neglected/*.

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3(4–5)**:993–1022 DOI 10.1162/jmlr.2003.3.4-5.993.

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5(1)**:135–146 DOI 10.1162/tacl_a_00051.

Business Today. 2020. Pakistan reports 730 active coronavirus cases; shuts all international flights. *Available at https://www.businesstoday.in/latest/economy-politics/story/pakistan-reports-730-active-coronavirus-cases-shuts-all-international-flights-252741-2020-03-22*.

Chaudhary S, Naaz S. 2017. Use of big data in computational epidemiology for public health surveillance. In: *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*. 150–155.

Choi S, Lee J, Kang M-G, Min H, Chang Y-S, Yoon S. 2017. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods* **129(19)**:50–59 DOI 10.1016/j.ymeth.2017.07.027.

Dawn. 2020. Over 450 held across Sindh on first day of lockdown. *Available at https://www.dawn.com/news/1543279/over-450-held-across-sindh-on-first-day-of-lockdown*.

Dawn. 2021. Pakistan secures 17 million doses of AstraZeneca. *Available at https://www.dawn.com/news/1604617*.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint* DOI 10.48550/arXiv.1810.04805.

Dubey AD. 2020. Twitter sentiment analysis during COVID-19 outbreak. *SSRN Electronic Journal* **8(2)** DOI 10.2139/ssrn.3572023.

Elder E, Johnston ANB, Wallis M, Crilly J. 2020. The demoralisation of nurses and medical doctors working in the emergency department: a qualitative descriptive study. *International Emergency Nursing* **52(8)**:100841 DOI 10.1016/j.ienj.2020.100841.

Erku DA, Belachew SA, Abrha S, Sinnollareddy M, Thomas J, Steadman KJ, Tesfaye WH. 2021. When fear and misinformation go viral: pharmacists' role in deterring medication misinformation during the 'infodemic' surrounding COVID-19. *Research in Social and Administrative Pharmacy* **17(1)**:1954–1963 DOI 10.1016/j.sapharm.2020.04.032.

Garcia K, Berton L. 2021. Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* **101(1)**:107057 DOI 10.1016/j.asoc.2020.107057.

Geo News. 2021. Pakistan to get COVID-19 vaccine by end of this month: parliamentary secretary. *Available at https://www.geo.tv/latest/328697-pakistan-to-get-covid-19-vaccine-by-end-of-this-month-parliamentary-secretary*.

Geo News. 2020a. Coronavirus updates, March 31: latest news on the coronavirus pandemic from Pakistan and around the world. *Available at https://www.geo.tv/latest/280043-coronavirus-updates-march-31-latest-news-on-the-coronavirus-pandemic-from-pakistan-and-around-the-world*.

Geo News. 2020b. Coronavirus vaccine: Pakistan to purchase 1.2m doses from China's Sinopharm. *Available at https://www.geo.tv/latest/327360-coronavirus-vaccine-pakistan-to-purchase-12m-doses-from-chinas-sinopharm*.

Goldschmidt K. 2020. The COVID-19 pandemic: technology use to support the wellbeing of children. *Journal of Pediatric Nursing* **53**:88–90 DOI 10.1016/j.pedn.2020.04.013.

Gupta RK, Yang Y. 2018. Crystalfeel at semeval-2018 task 1: understanding and detecting emotion intensity using affective lexicons. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. 256–263.

Hamzah FAB, Lau CH, Nazri H, Ligot DV, Lee G, Tan CL, Shaib MKBM, Zaidon UHB, Abdullah AB, Chung MH, Ong CH, Chew PY. 2020. CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ* **1(4)**:32 DOI 10.2471/BLT.20.255695.

Hutto C, Gilbert E. 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. 216–225. *Available at https://ojs.aaai.org/index.php/ICWSM/article/view/14550*.

Islam AN, Laato S, Talukder S, Sutinen E. 2020. Misinformation sharing and social media fatigue during COVID-19: an affordance and cognitive load perspective. *Technological Forecasting and Social Change* **159(1)**:120201 DOI 10.1016/j.techfore.2020.120201.

Islam T, Pitafi AH, Arya V, Wang Y, Akhtar N, Mubarik S, Xiaobei L. 2021. Panic buying in the COVID-19 pandemic: a multi-country examination. *Journal of Retailing and Consumer Services* **59**:102357 DOI 10.1016/j.jretconser.2020.102357.

Jabeen A, Ansari JA, Ikram A, Abbasi SH, Khan MA, Rathore TR, Safdar RM. 2020. Comparison of actions taken by Pakistan, United Arab Emirates and vietnam for COVID-19 prevention and control. *Global Biosecurity* **2**.

Jadeja G. 2022. Labeled twitter dataset. *Available at https://github.com/gauravjadeja3/Labeled-COVID-19-Twitter-Dataset*.

Joshi V, Peters M, Hopkins M. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. *ArXiv preprint* DOI 10.48550/arXiv.1805.06556.

Kaleej Times. 2020. Coronavirus: Pakistan suspends international flights for 2 weeks. *Available at https://www.khaleejtimes.com/coronavirus/coronavirus-pakistan-suspends-international-flights-for-2-weeks*.

Killick R, Fearnhead P, Eckley IA. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107(500)**:1590–1598 DOI 10.1080/01621459.2012.737745.

Kumar V. 2022. Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model. *Scientific Reports* **12(1)**:1–14 DOI 10.1038/s41598-022-05974-6.

Liu S, Liu J. 2021. Public attitudes toward COVID-19 vaccines on English-language Twitter: a sentiment analysis. *Vaccine* **39(39)**:5499–5505 DOI 10.1016/j.vaccine.2021.08.058.

Machuca CR, Gallardo C, Toasa RM. 2021. Twitter sentiment analysis on coronavirus: machine learning approach. *Journal of Physics Conference Series* **1828(1)**:012104 DOI 10.1088/1742-6596/1828/1/012104.

Mohammad SM, Turney PD. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* **29(3)**:436–465 DOI 10.1111/j.1467-8640.2012.00460.x.

OyeYeah. 2020. Sindh govt. restricts movement between 8am–5pm as coronavirus cases surge to 440. *Available at https://www.oyeyeah.com/news/sindh-restricts-movement-between-8-am-5-pm/*.

Pan X, Ojcius DM, Gao T, Li Z, Pan C, Pan C. 2020. Lessons learned from the 2019-nCoV epidemic on prevention of future infectious diseases. *Microbes and Infection* **22(2)**:86–91 DOI 10.1016/j.micinf.2020.02.004.

Pennington J, Socher R, Manning CD. 2014. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

Plutchik R. 1980. Chapter 1-A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H, eds. *Theories of Emotion*. Cambridge: Academic Press, 3–33.

Radovic M, Ghalwash M, Filipovic N, Obradovic Z. 2017. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **18(1)**:1–14 DOI 10.1186/s12859-016-1423-9.

Reimers N, Gurevych I. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. *ArXiv preprint* DOI 10.48550/arXiv.1908.10084.

Ridhwan KM, Hargreaves CA. 2021. Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights* **1(2)**:100021 DOI 10.1016/j.jjimei.2021.100021.

Samuel J, Ali GGMN, Rahman MM, Esawi EK, Samuel Y. 2020. Covid-19 public sentiment insights and machine learning for tweets classification. *Information-an International Interdisciplinary Journal* **11(6)**:314 DOI 10.3390/info11060314.

Singh M, Jakhar AK, Pandey S. 2021. Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining* **11(1)**:1–11 DOI 10.1007/s13278-021-00737-z.

Singh R, Singh R, Bhatia A. 2018. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. *International Journal of Advanced Science and Research* **3(2)**:19–24 DOI 10.14419/ijet.v7i2.6.10060.

Soliman HASA, Tabak F. 2020. Deep learning framework for RDF and knowledge graphs using fuzzy maps to support medical decision. *Journal of International Research in Medical and Pharmaceutical Sciences* **14(3)**:92–97.

Statesman T. 2020. Coronavirus pandemic: Pakistan to extend lockdown for 2 more weeks as death toll reaches 31. *Available at https://www.thestatesman.com/world/coronavirus-pandemic-pakistan-extend-lockdown-2-weeks-death-toll-reaches-31-1502872742.html*.

Surano FV, Porfiri M, Rizzo A. 2021. Analysis of lockdown perception in the United States during the COVID-19 pandemic. *The European Physical Journal Special Topics* 1–9 DOI 10.1140/epjs/s11734-021-00265-z.

Tan H, Peng S-L, Zhu C-P, You Z, Miao M-C, Kuai S-G. 2021. Long-term Effects of the COVID-19 pandemic on public sentiments in mainland China: sentiment analysis of social media posts. *Journal of Medical Internet Research* **23(8)**:e29150 DOI 10.2196/29150.

The Express Tribune. 2021. Pakistan to get 17m Covid-19 vaccine doses from UK. *Available at https://tribune.com.pk/story/2287746/pakistan-to-get-17m-covid-19-vaccine-doses-from-uk*.

The Express Tribune. 2020a. Sindh decides to go into COVID-19 lock down. *Available at https://tribune.com.pk/story/2181138/1-sindh-govt-mulls-total-lockdown-provincial-covid-19-tally-crosses-350/*.

The Express Tribune. 2020b. Sindh extends closure of all educational institutions till March 13 amid virus fear. *Available at https://tribune.com.pk/story/2167367/1-coronavirus-threat-educational-institutions-sindh-remain-shut-till-march-13/*.

The Guardian. 2020. Pakistan suspends international flights for two weeks. *Available at https://www.theguardian.com/world/2020/mar/21/pakistan-suspends-international-flights-for-two-weeks-to-stem-covid-19-spread*.

The International News. 2020. Sindh govt extends closure of educational institutions till May 30 amid coronavirus fears. *Available at https://www.thenews.com.pk/latest/628126-govt-extends-educational-institutions-holidays-till-may-30-amid-coronavirus-fears*.

Vahia IV, Jeste DV, Reynolds CF. 2020. Older adults and the mental health effects of COVID-19. *JAMA* **324(22)**:2253–2254 DOI 10.1001/jama.2020.21753.

Velavan TP, Meyer CG. 2020. The COVID-19 epidemic. *Tropical Medicine & International Health* **25(3)**:278–280 DOI 10.1111/tmi.13383.

Villavicencio C, Macrohon JJ, Inbaraj XA, Jeng J-H, Hsieh J-G. 2021. Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. *Information* **12(5)**:204 DOI 10.3390/info12050204.

Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, Abrego GH, Yuan S, Tar C, Sung Y-H, Strope B, Kurzweil R. 2019. Multilingual universal sentence encoder for semantic retrieval. *ArXiv preprint* DOI 10.48550/arXiv.1907.04307.

Zhang Q, Yi GY, Chen L-P, He W. 2021. Text mining and sentiment analysis of COVID-19 tweets. *ArXiv preprint* DOI 10.48550/arXiv.2106.15354.

Zhou X, Song Y, Jiang H, Wang Q, Qu Z, Zhou X, Jit M, Hou Z, Lin L. 2021. Comparison of public responses to containment measures during the initial outbreak and resurgence of COVID-19 in China: infodemiology study. *Journal of Medical Internet Research* **23(4)**:e26518 DOI 10.2196/26518.

PeerJ Computer Science

Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets

Ahmad Fathan Hidayatullah[1,2], Rosyzie Anna Apong[1], Daphne T.C. Lai[1] and Atika Qazi[3]

[1] School of Digital Science, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei
[2] Department of Informatics, Universitas Islam Indonesia, Sleman, Yogyakarta, Indonesia
[3] Centre for Lifelong Learning, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei

ABSTRACT

With the massive use of social media today, mixing between languages in social media text is prevalent. In linguistics, the phenomenon of mixing languages is known as code-mixing. The prevalence of code-mixing exposes various concerns and challenges in natural language processing (NLP), including language identification (LID) tasks. This study presents a word-level language identification model for code-mixed Indonesian, Javanese, and English tweets. First, we introduce a code-mixed corpus for Indonesian-Javanese-English language identification (IJELID). To ensure reliable dataset annotation, we provide full details of the data collection and annotation standards construction procedures. Some challenges encountered during corpus creation are also discussed in this paper. Then, we investigate several strategies for developing code-mixed language identification models, such as fine-tuning BERT, BLSTM-based, and CRF. Our results show that fine-tuned IndoBERTweet models can identify languages better than the other techniques. This is the result of BERT's ability to understand each word's context from the given text sequence. Finally, we show that sub-word language representation in BERT models can provide a reliable model for identifying languages in code-mixed texts.

Submitted 25 May 2022
Accepted 6 March 2023
Published 22 June 2023

Corresponding author
Ahmad Fathan Hidayatullah,
fathan@uii.ac.id,
21h2501@ubd.edu.bn

Academic editor
Lexing Xie

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.1312**

Copyright
2023 Hidayatullah et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Computational Linguistics, Data Mining and Machine Learning, Natural Language and Speech, Network Science and Online Social Networks, Text Mining
Keywords Code-mixing, Language identification, Indonesian, Javanese, English, Twitter, BERT

INTRODUCTION

Today, mixing languages is prevalent in daily communication, especially in informal situations, such as texting posts on social media. In linguistics, combining two or more languages within an utterance of speech or text is called code-mixing (*Hoffmann, 2014*; *Ritchie & Bhatia, 2012*). Mixing languages is particularly common in regions where people are natively multilingual.

Indonesia is one of the world's most multilingual countries, with over 700 local spoken languages (*Aji et al., 2022*). More than 198 million and 84 million people speak Indonesian and Javanese, respectively (*Eberhard, Simons & Fennig, 2021*). Hence, mixing Indonesian and Javanese in an utterance is common in Indonesia, especially among the Javanese people. Besides, exposure to English from social media and school makes Indonesians mix their

languages with English (*Rizal & Stymne, 2020*). As a result, mixing Indonesian, Javanese, and English in daily conversation becomes the most prevalent language combination in Indonesian societyt (*Yulianti et al., 2021*).

The following is an example of a code-mixed sentence containing Indonesian, Javanese, and English words:

"*Aku udah coba ngedownload tapi error, tulung aku diewangi downloadke panduane!*"

(*English:* ***I have tried to download but error, please help me download the guideline!***).

The sentence contains the following language compositions: Indonesian *(aku, udah, coba, tapi),* mix Indonesian-English *(ngedownload)*, English *(error)*, Javanese *(tulung, aku, diewangi)*, mix Javanese-English *(downloadke)*, and mix Indonesian-Javanese *(panduane)*. In the above sentence, the mixing of languages occurs not only within the sentence but also within the word. For example, the word '*ngedownload*' consists of '*nge-*' (informal Indonesian prefix) and '*download*' (English). The word '*downloadke*' consists of '*download*' (English) and '*-ke*' (Javanese suffix). The word '*panduane*' consists of '*panduan*' (Indonesian) and '*-e*' (Javanese suffix).

To analyze code-mixed text, a language identification (LID) task is often used as part of the pre-processing step (*Hidayatullah et al., 2022*). LID is critical for some subsequent natural language processing tasks in code-mixed documents (*Gundapu & Mamidi, 2018*). Applying LID in the code-mixed text has become a foundation work of various NLP systems, including sentiment analysis (*Ansari & Govilkar, 2018*; *Mahata, Das & Bandyopadhyay, 2021*), translation (*Barik, Mahendra & Adriani, 2019*; *Mahata et al., 2019*), and emotion classification (*Yulianti et al., 2021*). The absence of LID in pre-processing tasks can affect those NLP systems. For example, if the language is not accurately identified, a code-mixed sentence will produce an inaccurate translation. In another case, an offensive content identification system may produce incorrect results if the words in a sentence are not correctly identified (*Singh, Sen & Kumaraguru, 2018*).

However, most existing NLP systems are designed to process a single language at once (*Sabty et al., 2021*). The number of NLP systems that can process multiple languages per sentential unit is restricted (*Nguyen et al., 2021*). The traditional language identification systems fail to detect languages correctly from mixed language texts (*Kalita & Saharia, 2018*). Processing multiple languages within a sentence requires additional processing tasks compared to monolingual texts due to various language combinations such as sentence, clause, word, and sub-word levels (*Mave, Maharjan & Solorio, 2018*). Detecting language from code-mixed text using a traditional approach like dictionary lookup is no longer applicable. The dictionary approach produces poor results due to spelling inconsistencies and the loss of word context (*Ansari & Govilkar, 2018*).

On the other hand, the availability of annotated code-mixed data, including Indonesian and Javanese data, remains limited. Even though Indonesian and Javanese have many speakers, only a few studies have addressed the code-mixing phenomenon in the Indonesian language (*Adilazuarda et al., 2022*; *Winata et al., 2022*). In comparison to the languages spoken in Europe, the existence of Indonesian and Javanese languages in NLP research is relatively understudied (*Aji et al., 2022*).

Considering the problems above, this study makes the following contributions:

1. We develop an annotated code-mixed corpus for Indonesian-Javanese-English language identification (IJELID) from Twitter data.
2. To build the language identification model, we apply a fine-tuning strategy based on pre-trained Bidirectional Encoder Representations from Transformers (BERT) models introduced by *Devlin et al. (2019)*. BERT has shown outstanding performance when fine-tuned for any downstream NLP tasks (*Ulčar & Robnik-Šikonja, 2020*). We also seek to use several strategies using CRF and BLSTM-based models.

RELATED WORK

Code-mixed data availability for language identification

In this study, we collect some papers focused on conducting language identification for code-mixed text. As a result, we found 17 related studies published between 2016 and 2022. During that period year, we identify 14 code-mixed datasets such as Manipuri-English (*Lamabam & Chakma, 2016*), Konkani-English (*Phadte & Wagh, 2017*), Telugu-English (*Gundapu & Mamidi, 2018*), Bengali-English (*Jamatia, Das & Gambäck, 2018*; *Mandal & Singh, 2018*), Hindi-English (*Ansari et al., 2021*; *Jamatia, Das & Gambäck, 2018*; *Mandal & Singh, 2018*; *Shekhar, Sharma & Beg, 2020*), Bengali-Hindi-English (*Jamatia, Das & Gambäck, 2018*), Turkish-English (*Yirmibeşoğlu & Eryiğit, 2018*), Indonesian-English (*Barik, Mahendra & Adriani, 2019*; *Yulianti et al., 2021*), Sinhala-English (*Smith & Thayasivam, 2019*), Arabic-English (*Sabty et al., 2021*), English-Assamese-Hindi-Bengali (*Sarma, Singh & Goswami, 2022*), Telugu-English (*Kusampudi, Chaluvadi & Mamidi, 2021*), Malayalam-English (*Thara & Poornachandran, 2021*), and Kannada-English (*Shashirekha et al., 2022*; *Tonja et al., 2022*).

From those 14 datasets, five studies utilized Indonesian-related code-mixed datasets. *Barik, Mahendra & Adriani (2019)* introduced code-mixed Indonesian-English data from Twitter for the text normalization task. *Yulianti et al. (2021)* used the dataset created by *Barik, Mahendra & Adriani (2019)* to see the impact of code-mixed normalization on the emotion classification task. In *Yulianti et al. (2021)*, they introduced new feature sets to improve the performance of the language identification task. *Suciati & Budi (2019)* created a review dataset containing mixed Indonesian-English. They gathered the review data from a culinary website for an aspect-based opinion mining task. *Arianto & Budi (2020)* also developed a code-mixed Indonesian-English dataset for aspect-based sentiment analysis. The dataset was collected from Google Maps reviews. *Tho et al. (2021)* proposed a code-mixed Indonesian-Javanese corpus using the Twitter dataset for sentiment analysis.

Among the five previous studies, two papers (*Barik, Mahendra & Adriani, 2019*; *Yulianti et al., 2021*) applied language identification in their research. The remaining studies focused on opinion mining and sentiment analysis. All five studies above provided bilingual code-mixed languages, namely Indonesian-English and Indonesian-Javanese. There have been no studies focusing on trilingual code-mixed language data, particularly code-mixed Indonesian-Javanese-English. The existing studies do not concentrate on the language identification task. Instead, they aimed to solve problems in sentiment analysis, emotion classification, and translation. Since no dataset is available for code-mixed

Indonesian-Javanese-English, it is necessary to develop a dataset for such mixed languages. The availability of a code-mixed Indonesian-Javanese-English dataset can help improve NLP research for related Indonesian languages.

Language identification for code-mixed text

Many studies have developed language identification models for code-mixed text using various techniques. CRF has been utilized in many research due to its simplicity and impressive performance in building a code-mixed language identification model (*Gundapu & Mamidi, 2018*). *Lamabam & Chakma (2016)* applied a trigram-based model using J48 and CRF for code-mixed English-Manipuri data. Their experiments revealed that the CRF with the first three and last three-character features gave the best result with an F1 score of 90%. *Phadte & Wagh (2017)* applied support vector machine (SVM), random forest, and CRF to build a word-level language identification system code-mixed Konkani-English text. In their study, CRF outperformed the other methods with an accuracy of 97%. *Gundapu & Mamidi (2018)* worked on English-Telugu code-mixed data by comparing four algorithms, naïve Bayes, random forest, hidden Markov model, and CRF. CRF achieved the best performance with an F1 score of 91% and an accuracy of 91.28%. *Yirmibeşoğlu & Eryiğit (2018)* presented a word-level language identification system for code-mixed Turkish-English using CRF and n-gram character level as a feature. The CRF could give a promising performance with an F1 score macro of 94.5%. *Smith & Thayasivam (2019)* built language identification models for code-mixed Sinhala-English. In their study, CRF outperformed SVM, LSTM, KNN, and random forest, with an overall F1 score of 94%. A study by *Barik, Mahendra & Adriani (2019)* conducted language identification using mixed Indonesian-English data from Twitter using CRF and obtained an F1 score of 89.58%. *Yulianti et al. (2021)* used the same dataset as *Barik, Mahendra & Adriani (2019)* to develop a language identification system. They applied various techniques, such as CRF, CRF++, CNN, CNN+CRF, and RNN. The results showed that CRF++ performed the best, with an F1 score of 93.98%.

Mandal & Singh (2018) proposed a multichannel neural network-based method for code-mixed Bengali-English and Hindi-English. The multichannel module combines convolutional neural network (CNN) and long-short term memory (LSTM). In addition, they added a bidirectional-LSTM-CRF module to capture the context of the input text. *Jamatia, Das & Gambäck (2019)* presented a Hindi-Bengali-English code-mixed language identification using LSTM and BLSTM. Based on their experiments, the BLSTM performed better than the LSTM, with an F1 score of 87.07%. *Shekhar, Sharma & Beg (2020)* built a code-mixed language identification model for Hindi-English using Bidirectional LSTM (BLSTM). Their study revealed that BLSTM achieved an F1 score of 93.97%.

Sabty et al. (2021) utilized a character BLSTM and segmental recurrent neural network (SegRNN) for the Arabic-English text language identification system. The SegRNN achieved the best results in their experiments with an F1 score of 94.84%. *Sarma, Singh & Goswami (2022)* studied a word-level language identification system for Assamese, Bengali, Hindi, and English code-mixed data utilizing CNN and BLSTM with pre-trained word embeddings. Their experiments indicated that the CNN performed better than the

BLSTM, with an F1 score of 90.79%. *Kusampudi, Chaluvadi & Mamidi (2021)* created a code-mixed Telugu-English corpus. They conducted word-level language identification using BLSTM-CRF with an accuracy of 99.32%.

Thara & Poornachandran (2021) developed a large-scale Malayalam-English dataset for code-mixed language identification using BERT-based models, such as XLM-RoBERTa, ELECTRA, CamemBERT, and DistilBERT. Compared to the other BERT models, ELECTRA obtained the highest F1 score of 99.33% and an accuracy of 99.41%. *Ansari et al. (2021)* leveraged BERT base and RoBERTa models for code-mixed Hindi-English language identification. Their experiments showed that pre-training and fine-tuning using code-mixed Hindi-English-Urdu text yielded an F1-score of 84%, higher than the pre-trained monolingual models. *Shashirekha et al. (2022)* developed a code-mixed Kannada-English dataset called CoLI-Kenglish. In their study, three different approaches were applied, such as machine learning-based (CoLI-vectors and CoLI-ngram), deep learning-based (CoLI-BiLSTM), and transfer learning-based (CoLI-BiLSTM). CoLI-ngrams using morphological features outperformed all other models with an average macro F1-score of 64%. *Tonja et al. (2022)* combined LSTM and BERT for code-mixed Kannada-English language identification. Their proposed model gained a weighted F1 score of 84% and a macro F1 score of 61%.

To summarize, NLP researchers have applied various techniques for code-mixed language identification tasks. CRF has shown a satisfying performance among traditional machine learning techniques. LSTM-based architectures have demonstrated promising results in several studies. The emergence of the transformer (*Vaswani et al., 2017*) has improved NLP research. Furthermore, transformer-based methods, such as BERT and its variants, have become a breakthrough and state-of-the-art in solving NLP problems (*Wolf et al., 2020*). BERT-based models have shown remarkable performance in the code-mixed language identification task and outperformed traditional neural network techniques, such as RNN and CNN (*Thara & Poornachandran, 2021*).

CORPUS CREATION

Data collection and pre-processing

We collect 15K tweets in several batches from December 2021 to July 2022. In this work, we list some code-mixed Indonesian, Javanese, and English keywords to obtain the tweets to ensure that the retrieved tweets are code-mixed. In the pre-processing tasks, we first filter the tweets by removing duplicates. Subsequently, we replace user mentions, URLs, and hashtags with @user, httpurl, and #hashtag. Finally, we convert all words into lowercase. In our dataset, we notice some tweets containing languages other than Indonesian, Javanese, and English, such as Malay, Sundanese, Arabic, and Korean. The occurrence of Malay and Sundanese in the retrieved tweets is reasonable since both languages have similarities to Indonesian and Javanese. In addition, Arabic and Korean scripts exist in our dataset because people sometimes add both scripts to their tweets. In this case, we keep those languages in our dataset.

Annotation guidelines

To develop the dataset, we employ two annotators who understand Indonesian, Javanese, and English. We cover standard and non-standard language labeling forms in the annotation process. Since we work on social media text, the characteristics of the data are different from the standard text. The texts consist of many informal forms, such as slang words, non-standard spellings, mixed letters with numbers, abbreviated words, and lengthening words. Hence, this study proposes an annotation guideline to cope with those informal texts. The following is an annotation guideline to produce a standard annotation result:

1. This study introduces seven labels to annotate tokens from the dataset, namely: ID (Indonesian), JV (Javanese), EN (English), MIX_ID_EN (mixed Indonesian-English), MIX_ID_JV (mixed Indonesian-Javanese), MIX_JV_EN (mixed Javanese-English), and OTH (Other).
2. The ID label is used to annotate Indonesian words, including both standard and non-standard forms of Indonesian words.
3. The EN label is used to annotate English words, including both standard and non-standard forms of English words.
4. The JV label is used to annotate Javanese words, including both standard and non-standard Javanese words.
5. The MIX_ID_EN label is used to annotate words containing mixed Indonesian-English. For example:
   - *diprint (printed)* → *di- (Indonesian prefix) + print (English).*
   - *filenya (the file)* → *file (English) + -nya (Indonesian suffix).*
6. The MIX_ID_JV label is used to annotate words containing mixed Indonesian-Javanese. For instance:
   - *bajune (the clothes)* → *baju (Indonesian) + ne (Javanese suffix).*
   - *penjuale (the seller)* → *penjual (Indonesian) + e (Javanese suffix).*
7. The MIX_JV_EN label is used to annotate words containing mixed Javanese-English. For example:
   - *endinge (the ending)* → *ending (English) + -e (Javanese suffix).*
   - *ngemailke (send email)* → *ng- (Javanese prefix) + email (English) + -ke (Javanese prefix).*
8. The OTH (Other) label is used to annotate the following:
   - symbols, punctuations, and number,
   - unknown words, words other than Indonesian, Javanese, and English,
   - Twitter entities: username or mention,
   - URL, emoji, emoticons, date, time, and currency,
   - laugh expressions *(haha, xixixi, wkwkwk)* and sad expressions *(huhuhu, hiks).*

9. If a word *(w)* exists in more than one language, it will be annotated as a particular language (ID, JV, or EN), depending on the context.
   - If *w* is preceded and followed by Javanese words, *w* is annotated as JV.
   - If *w* is preceded and followed by Indonesian words, *w* is annotated as ID.
   - If *w* is preceded and followed by English words, *w* is annotated as EN.
10. Foreign or borrowed words that exist in all languages are identified depending on the context of tweets or sentences.
11. A named entity followed by a possessive pronoun from a particular language can be labeled as ID or JV, depending on the context.
    - *Macbookku (My Macbook)* → *Macbook (named entity) + -ku (Indonesian or Javanese suffix).*
    - *Macbooknya (The Macbook)* → *Macbook (named entity) + -nya (Indonesian suffix).*

Corpus creation challenges

This sub-section highlights some challenges encountered during the data creation process. Annotating language in Twitter data is not as simple as in formal text data due to the non-standard language forms. In this study, we define the non-standard language forms by categorizing into the following: mixed letters and numbers, slang, abbreviated words and acronyms, expressive lengthening, and English words written in Indonesian spelling style (*Hidayatullah, 2015*). Table 1 presents examples of non-standard form words in Indonesian, Javanese, and English.

Another challenge in the annotation stage is determining words recognized in more than one language. For instance,

(1) *Guys, pertandingan **final** nanti malam jam berapa? Thanks*
English: *Guys, what time is the final match tonight? Thanks*

(2) *Finally, Liverpool menang di babak quarter **final** UCL.*
English: *Finally, Liverpool won in the UCL quarter-finals.*

From the two examples above, the word *'final'* exists in Indonesian and English languages. Therefore, we must look at the surrounding words and the sentence's whole context to determine the language label for the word *'final'*. In the first example, the word *'final'* is labeled as ID because the previous and the following words are Indonesian. In the second example, the word *'final'* is labeled as EN because the term *'quarter final'* is an English phrase.

Moreover, identifying intra-word code-mixing is also challenging due to the prefixes or suffixes that exist in the two languages. For example, the word *'disave'* (saved) consists of *'di-'* (a prefix) followed by the English word *'save'*. Another example is *'smartphoneku'* (my smartphone), which contains *'smartphone'* and the suffix *'-ku'*. The prefix *'di-'* in the word *'disave'* and the suffix *'-ku'* in *'smartphoneku'* exist in Indonesian and Javanese. Therefore, the word *'disave'* and *'smartphoneku'* can be annotated as MIX_ID_EN or MIX_JV_EN, depending on the context of the sentence. To determine the correct label, the annotators should consider the previous words, the following words, and the whole words of the sentence.

Table 1 Examples of the non-standard form words in Twitter data.

No	Type	Language		
		Indonesian	Javanese	English
1	Mixed letters and numbers	•'*se7*' stands for '*setuju*' (to agree)	• '*siji2*' stands for '*siji-siji*' (one by one)	•'*b4*' (before)
		•'*anak2*' stands for '*anak-anak*' (kids)		•'*ni8t*' (night)
2	Slangs	•'*santuy*' stands for '*santai*' (relaxed)	•'*ngunu*' stands for '*ngono*' (like that)	•'*epic*' (awesome)
		•'*nobar*' stands for '*nonton bareng*' (watch together)	•'*bonek*' stands for '*bondo nekat*' (reckless)	•'*noob*' (newbie)
				•'*menfess*' (mention confess)
				•'*crashy*' (crazy and trashy)
3	Abbreviated words or acronym	•'*dgn*' stands for '*dengan*' (with)	•'*kbh*' stands for '*kabeh*' (all)	•'*idk*' *(I don't know)*
		•'*slkh*' stands for '*sekolah*' (school)	•'*mtrnwn*' stands for '*matur nuwun*' (thank you)	•'*dm*' *(direct message)*
		•'*SD*' stands for '*sekolah dasar*' (elementary school)	•'*lur*' stands for '*sedulur*' (brother or sister)	•'*omg*' *(oh my god)*
				•'*thx*' *(thanks)*
				•'*ur*' (your)
4	Expressive lengthening	• '*senaaang*' stands for '*senang*' (happy)	• '*kaaabeeeh*' stands for '*kabeh*' (all)	• '*gooooood*' (good)
5	English words written in Indonesian spelling style	–	–	•'*n*' (and)
				•'*plis*' (please)
				•'*wiken*' (weekend)
				•*donlod* (download)
				•'*gud*' (good)
				•'*selow*' (slow)

Inter-annotator agreement

This study uses Cohen's kappa (*Cohen, 1960*) to quantify the agreement between two annotators. The following is the equation for calculating Cohen's kappa (K):

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}. \quad (1)$$

$Pr(a)$ is the frequency with which two annotators assigned the same label. $Pr(a)$ is obtained by calculating all agreed labels divided by the total data. $Pr(e)$ is the probability of agreement when the annotators see the observed data randomly. The $Pr(e)$ is calculated by summing the probability when both annotators randomly select the first label and the probability when both annotators select the second label. Cohen's kappa value (K) spans from 1 to -1, indicating that annotators select different labels for each sample. A value of 0 indicates that the annotators agreed precisely as frequently as they would if they were both guessing randomly. Therefore, the closer the K value to 1, the better the dataset.

Code-mixing index

After getting the distribution of each label from the dataset, we measure the complexity of code-mixing in the text using the code-mixing index (CMI) (*Gambäck & Das, 2014*). The CMI value is calculated using the following formula:

$$CMI = \begin{cases} 100 \times \left[1 - \dfrac{\max(w_{i:i=1,2,\ldots,n})}{n-u}\right], & n > u \\ 0, & n = u. \end{cases} \tag{2}$$

In Eq. (2), w_i is the number of words of the language i, $\max(w_i)$ represents the number of words of the prominent language, n is the total number of tokens, and u represents the number of language independent tokens (such as named entities, abbreviations, mentions, and hashtags). For monolingual utterances, the CMI score equals to 0 (zero), since the $\max(w_i) = n - u$. A low CMI score implies monolingualism in the text, whereas a high CMI score indicates code-mixing.

LANGUAGE IDENTIFICATION MODEL

Conditional random fields (CRF)

CRF is a discriminative model for sequential data labelling based on a conditional distribution approach (*Lafferty, McCallum & Pereira, 2001*). CRF employs contextual information from previous labels, thus increasing the amount of information for the model to generate an accurate label prediction. Let X be an input sentence containing sequence of words $(x_1, x_2, x_3, \ldots, x_n)$ and $Y = (y_1, y_2, y_3, \ldots, y_n)$ is the label prediction of such words. The CRF represents the likelihood of predicting the output Y given the input X as a conditional distribution $p(y|x)$. The calculation of a linear-chain CRF is shown by Eq. (3) (*Sutton & McCallum, 2012*):

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right\}. \tag{3}$$

In Eq. (3), $Z(x)$ is a normalization factor function obtained using the following formula:

$$Z(x) = \sum_{y} \prod_{t=1}^{T} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right\}. \tag{4}$$

In Eqs. (3) and (4), $\lambda = \{\lambda_k\}$ be a parameter vector weight estimated from the training set, and $F = \{f_k(y_t, y_{t-1}, x_t)\}_{k=1}^{K}$ be a set of feature functions. T is the number of time step indexed by t. K is the number of features and k indexes the feature function f_k and weight λ_k.

BLSTM-based architecture

Input representation for BLSTM-based architecture

1. Word-level representation

Word-level representation is a real-valued vector representing a word to capture the semantic relationship between words. Using word embedding, words with a closer

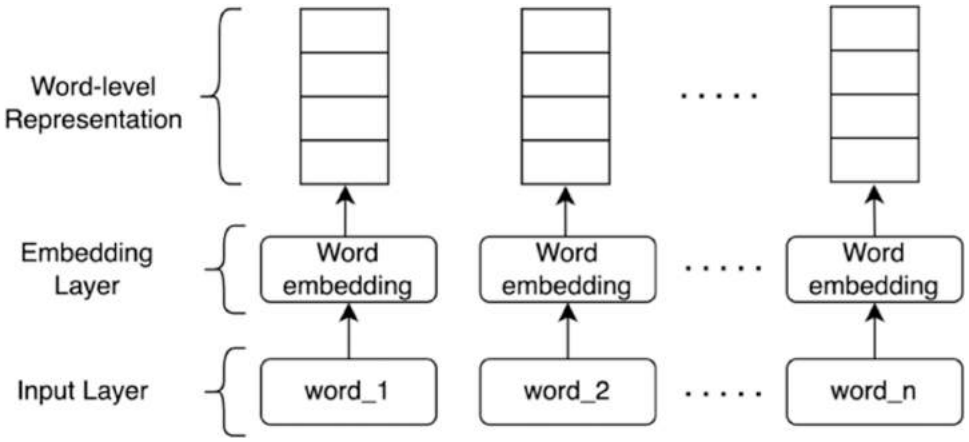


Figure 1 **Word-level representation.**
Full-size DOI: 10.7717/peerjcs.1312/fig-1

meaning will have a similar vector representation. The embedding layer carries out the transformation from words to their corresponding vector representations. The embedding layer receives sentences in the form of a sequence of tokens. Each token is then transformed into a word vector with a fixed size by mapping the index of such a token. Figure 1 illustrates the word-level representation model.

2. Character-level representation

Character-level representation aims to capture morphological features of the words by processing the character that composes words (*Mave, Maharjan & Solorio, 2018*). Using character-level representation can help alleviate the out-of-vocabulary (OOV) problems in the text data (*Joshi & Joshi, 2020*). The results of character representation are used to augment the word vector representation before being processed through the classification layer.

This study employs CNN and LSTM to train the character-level representation. In character-based CNN representation, a word is decomposed into a sequence of characters. The character inputs are passed into the character embedding layer. The embedding layer outputs are transmitted to the convolutional layer, which produces local features by applying a convolutional filter across a sliding n-character window. Subsequently, the max-pooling layer takes the maximum value over each dimension to represent a particular word. As for the character LSTM representation, we put a single LSTM layer on top of the embedding layer. The embedding outputs pass the character embedding vector via forward and backward LSTMs and combine each output to generate the encoding of the associated word. The character CNN and character LSTM representations are illustrated in Fig. 2.

Bidirectional long short-term memory (BLSTM)

BLSTM is an enhancement of LSTM networks that tries to increase model performance on sequence classification tasks (*Hidayatullah, Cahyaningtyas & Pamungkas, 2020*). By incorporating two distinct hidden layers, bidirectional LSTM (BLSTM) networks can access both preceding and subsequent contexts. These networks can efficiently capture

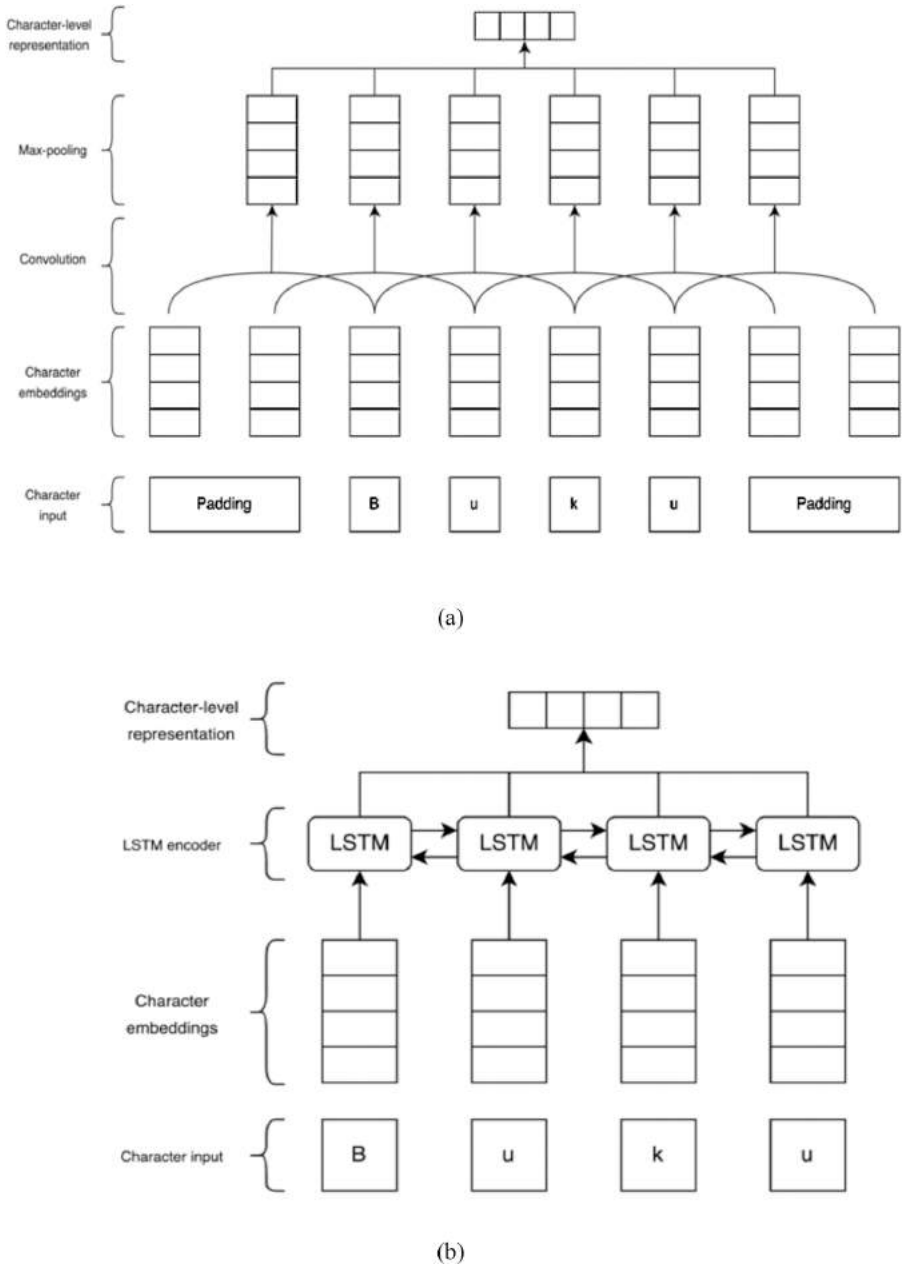


Figure 2 **(A) Character CNN representation. (B) Character LSTM representation.** These illustrations are adapted from *Reimers & Gurevych (2017)*.

Full-size DOI: 10.7717/peerjcs.1312/fig-2

long-distance relations in the sequence in both directions (*Mave, Maharjan & Solorio, 2018*). We initialize the embedding layer and feed the output sequence to the spatial dropout layer with a dropout. Subsequently, we add a single BLSTM layer and a recurrent sigmoid activation. BLSTM applies two LSTM networks by incorporating forward and backward hidden layers to capture previous and subsequent contexts (*Liu & Guo, 2019*).

Lastly, a dense layer with a Softmax activation function is added to predict the final word label.

BLSTM-CRF

BLSTM-CRF is a combination of deep learning and traditional machine learning approaches. The combination between BLSTM and CRF has been proven to produce good results in sequence tagging tasks (*Poostchi, Borzeshi & Piccardi, 2018*). A CRF layer can be added to the top layer of the BLSTM architecture to predict the label of the entire phrase at the same time (*Lafferty, McCallum & Pereira, 2001*). In the sequence labeling task, it is critical to consider the association between neighboring labels. BLSTM, on the other hand, does not generalize the connection between output labels (*Wintaka, Bijaksana & Asror, 2019*). That is due to the probability distribution of BLSTM being independent. The combination of BLSTM and CRF may efficiently exploit past and future input features through an LSTM layer and sentence-level tag information through a CRF layer (*Huang, Xu & Yu, 2015*). In this BLSTM-CRF architecture, we stack the following layers: input, embedding, SpatialDropout1D, BLSTM, dense, and CRF.

BLSTM + character LSTM

In this architecture, we replicate the previous model proposed by *Samih et al. (2016)*. This model receives two different inputs: word and character. Following that, the embedding layer processes the word and character inputs. A single LSTM layer is used to generate the character-level representations. These vector representations of characters are then concatenated with vector representations of words. The concatenated vector results then pass through SpatialDropout1D and a BLSTM layer with 100 hidden units. Finally, a dense layer is added to the BLSTM to predict the tag results.

BLSTM + character CNN

Similar to the BLSTM + character LSTM method, the input layer of this BLSTM + character CNN used word and character embeddings. In this architecture, the character-level representations are trained using CNN (*Zhang, Zhao & LeCun, 2015*). The character representations are obtained by passing the character input to a dropout layer followed by a Conv1D (kernel_size: 2, activation: ReLu) and a MaxPooling1D layer. The outputs are then flattened and fed into a dropout layer. A SpatialDropout1D layer receives the concatenated vectors between word and character representations. BLSTM proceeds with the output from the SpatialDropout1D and passes the results to the final dense layer to get the tag prediction.

BERT-based architecture

BERT input representation

BERT has a particular set of rules for representing the input text, namely sub-word representation. Sub-word representation is an alternative solution between word and character-based representations. BERT uses the WordPiece tokenization algorithm to create the sub-word representation (*Wu et al., 2016*). WordPiece starts by establishing an initial vocabulary composed of elementary units and then increases this vocabulary to the

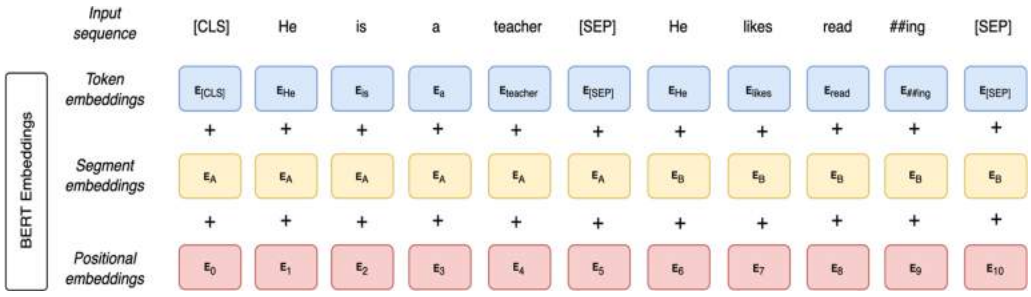


Figure 3 **BERT input representation.**
Full-size DOI: 10.7717/peerjcs.1312/fig-3

desired size. The vocabulary begins with characters from a single language. Then the most common character combinations in the vocabulary are added iteratively. WordPiece learns merged rules for the character pairs and finds the pair that maximizes the likelihood of the training data. Equation (5) is the formula to calculate the score for each pair:

$$\text{Score} = \frac{\text{frequency of pair}}{(\text{frequency of first unit} \times \text{frequency of second unit})}. \tag{5}$$

The score is calculated by dividing the frequency of the pair by the product of the frequencies of each of its components. The algorithm works by prioritizing the merging of pairs when each part occurs less frequently in the vocabulary. For example, the pair *'read'* and *'##ing'* will not be merged even though the token *'reading'* frequently appears in the vocabulary. This is because the pair *'read'* and *'##ing'* will probably frequently occur in many other words. A pair between *'re'* and *'##ad'* will likely be merged since *'re'* and *'##ad'* appear less frequently individually. Therefore, the token *'read'* is not split, while the token *'reading'* is separated into *'read'* and *'ing'*. This teaches the idea that the token *'reading'* is derived from *'read'* with slightly different meanings but the same origin.

As illustrated in Fig. 3, BERT input representation consists of three embeddings: token embeddings, segment embeddings, and position embeddings. In the token embeddings, two special tokens are added to each sentence. At the beginning of each sentence, a [CLS] token is added. Another special token is a [SEP] token which is located at the end of each sentence. The [SEP] token is added to separate between sentences. It is used as a learned segment embedding denoting a token as part of segment A or B. Segment embeddings are sentence numbers encoded in a vector. The model identifies whether a specific token belongs to sentence A or B in the segment embeddings. Position embeddings provide information regarding the word order in the input sequence. Finally, the BERT representation is obtained by summing those three embeddings.

BERT

Bidirectional encoder representations from transformers (BERT) is a language representation model built using the transformer-based technique developed by Google (*Devlin et al., 2019*). BERT is a transformer encoder stack capable of simultaneously reading a whole sequence of inputs. The BERT architecture is a deep bidirectional model, meaning that BERT takes information from both the left and right sides of the token's context during

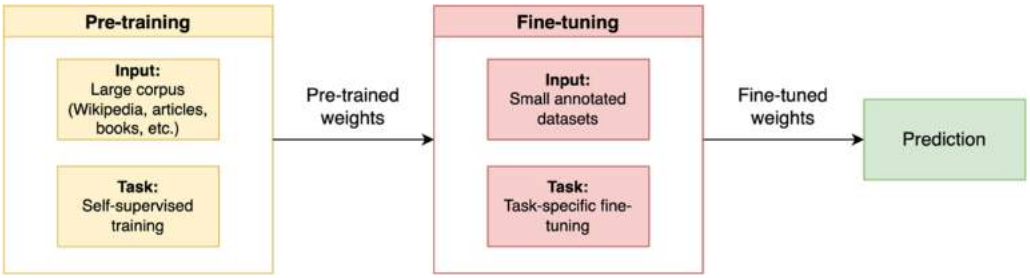


Figure 4 **Pre-training and fine-tuning BERT.**
Full-size DOI: 10.7717/peerjcs.1312/fig-4

the training task. BERT employs a self-multi-headed transformer-based attention model pre-trained on a large multilingual corpus (*Dowlagar & Mamidi, 2022*). BERT pre-trains the raw text data in a self-supervised way, which means there is no human annotation to generate inputs and labels from those texts.

BERT is pre-trained with two objectives, masked language modeling (MLM) and next sentence prediction (NSP). MLM randomly masks 15% of the words in the input text, then runs the whole masked sentence through the model to predict the masked words. For the NSP task, the model concatenates two masked sentences during pre-training and occasionally matches sentences adjacent to each other in the original text. The model then predicts whether the two sentences follow each other.

Fine-tuning BERT

As illustrated in Fig. 4, fine-tuning is done by leveraging a pre-trained model and then training it on a particular dataset suited to a specific task. The BERT model is first set with the pre-trained weight parameters. Next, all parameters are fine-tuned using annotated data from the downstream tasks. Ultimately, the fine-tuned weights are then used for the prediction task.

In this study, the fine-tuning tasks are performed by leveraging two existing pre-trained BERT models, namely multilingual BERT (mBERT) (https://huggingface.co/bert-base-multilingual-cased) and IndoBERTweet (https://huggingface.co/indolem/indobertweet-base-uncased). The first pre-trained model is created based on the multilingual BERT. Multilingual BERT is a masked language modeling (MLM) objective-trained model (*Devlin et al., 2019*). It is trained with a large Wikipedia corpus on top of 104 languages, including English, Indonesian, and Javanese. Another pre-trained model used to build the language identification model is IndoBERTweet. IndoBERTweet is a pre-trained domain-specific model using a large set of Indonesian Twitter data (*Koto, Lau & Baldwin, 2021*).

The illustration of fine-tuning BERT for the language identification task can be seen in Fig. 5. The [CLS] symbol and [SEP] are added at the beginning and the end of a single text sequence. Each token of the sequence and the contextual representation of each token are denoted by E and R, respectively. Following that, the BERT representation of each token is fed into dense layers. In the dense layers, the dense layer parameters are shared to get the label of each token.

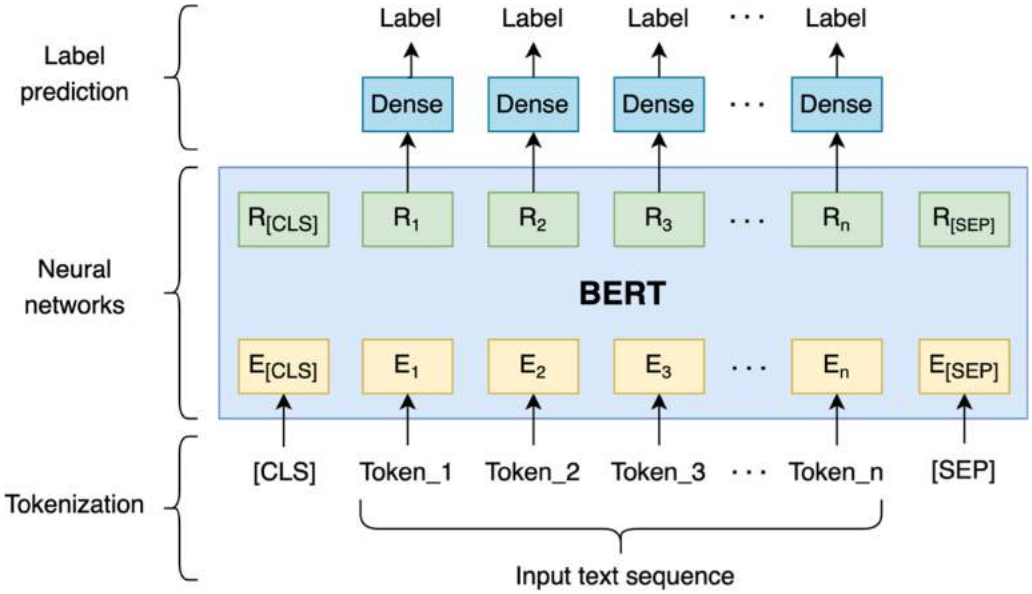


Figure 5 **Fine-tuning BERT for code-mixed language identification.**
Full-size DOI: 10.7717/peerjcs.1312/fig-5

Experimental setup

We conduct the experiments by splitting our dataset into training (6,170 tweets), validation (1,543 tweets), and testing (3,306 tweets). The training is performed using an 80 cores CPU, 250GB of RAM, and 4 GPUs (NVIDIA Tesla V100 SXM2). Before training, we apply hyperparameter tuning to get the optimal hyperparameters for each technique to maximize model performance.

For CRF training, the L-BFGS algorithm is utilized for gradient descent optimization and getting model parameters. We apply randomized search using *RandomizedSearchCV* (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html) to find the best parameter values for *L1* (Lasso) and *L2* (Ridge) regularization coefficients in the CRF algorithm. The hyperparameter search is done by applying 5-fold cross-validation. In addition, we set the number of parameter settings to 20.

In the BLSTM-based experiments, we use a grid search algorithm for hyperparameter tuning. First, we set several values to find the best values for learning rate, batch size, dropout, and the number of LSTM units. The best learning rate and dropout value for all models are 0.01 and 0.5, respectively. The batch size and LSTM units for BLSTM and BLSTM + character CNN architectures are 32 and 64, respectively. As for BLSTM-CRF and BLSTM + character LSTM, the batch size and LSTM units are 64 and 32, respectively. Then, the BLSTM-based models are trained using the Adam optimizer over 20 epochs. Table 2 provides the best hyperparameter values for each BLSTM-based model.

We also employ a grid search algorithm for fine-tuning BERT models. The hyperparameter search is conducted by setting categorical values for learning rate, per device training batch size, and per device evaluation batch size. The learning rate is set to 1e−4, 3e−4, 2e−5, 3e−5, and 5e−5. For the per-device train and evaluation batch size,

Table 2 Best hyperparameter values for each BLSTM-based model.

Model	Learning rate	Batch size	Dropout	LSTM units
BLSTM	0.01	32	0.5	64
BLSTM-CRF	0.01	64	0.5	32
BLSTM + character LSTM	0.01	64	0.5	32
BLSTM + character CNN	0.01	32	0.5	64

Table 3 Best hyperparameter values for each BERT-based model.

Model	Learning rate	Per device train batch size	Per device evaluation batch size
mBERT	3e−05	64	32
IndoBERTweet	2e−05	32	8

Table 4 Dataset information.

Number of tweets	11019
Number of tokens	273323
Number of unique tokens	44809
Average token length	4.63 characters
Average sentence length	24.8 tokens

we select the categorical values of 8, 16, 32, 64, and 128. Subsequently, the hyperparameter search is run by applying such defined values over four epochs and ten trials. The best hyperparameter values for each BERT-based model are presented in Table 3. Finally, we use the best hyperparameters with the Adam optimizer and five epochs for fine-tuning BERT training.

RESULTS AND DISCUSSION

Data annotation result

As described in Table 4, the dataset of this study consists of 11,019 tweets. We tokenize each tweet and annotate each token with the corresponding language label. As a result, our dataset comprises 273,323 tokens, 44,809 unique tokens, a tweet-length average of 24.8 tokens, and a token length average of 4.63 characters per token.

The Indonesian language dominates our dataset with 144,934 tokens (53%). The number of OTH (Other) is 62,254 tokens (22.8%), followed by Javanese and English with 29,835 tokens (10.9%) and 28,058 (10.3%), respectively. For the intra-word code-mixing data, Indonesian-English has the highest frequency among the intra-word code-mixing labels with 5,625 tokens (2.1%). In contrast, the number of intra-word code-mixing labels for Indonesian-Javanese and Javanese-English are 1,557 (0.6%) and 1,060 (0.4%), respectively. Table 5 shows the dataset statistics used in our study.

We evaluate the annotation results by measuring inter-annotator agreement based on the annotation results using Cohen's kappa. The agreement between the annotators

Table 5 Dataset statistics.

Label	Number of tokens	Percentage
Indonesian (ID)	144,934	53%
Javanese (JV)	29,835	10.9%
English (EN)	28,058	10.3%
Mixed Indonesian-English (MIX_ID_EN)	5,625	2.1%
Mixed Indonesian-Javanese (MIX_ID_JV)	1,557	0.6%
Mixed Javanese-English (MIX_JV_EN)	1,060	0.4%
Other (OTH)	62,254	22.8%
Total	273,323	100%

Table 6 Precision, recall, and F1 score results on the test set for each model.

Model	Precision		Recall		F1 score	
	Macro	Weighted	Macro	Weighted	Macro	Weighted
CRF	94.39	95.71	91.60	95.71	92.94	95.69
BLSTM	93.81	93.78	88.73	93.68	91.06	93.66
BLSTM-CRF	92.35	92.87	87.94	92.65	89.91	92.67
BLSTM + character LSTM	91.78	94.26	90.31	94.23	91	94.2
BLSTM + character CNN	93.51	93.69	89.04	93.67	91.11	93.66
mBERT	93.09	94.87	92.85	94.85	92.96	94.85
IndoBERTweet	93.21	95.64	93.86	95.63	93.53	95.63

attains a value of 0.9964. This result represents an almost perfect agreement between two annotators. As for the CMI, we gain a score of 38.05. It means that 38.05% of the overall non-neutral language tokens in the dataset are code-mixed.

Language identification results

In this sub-section, we provide the code-mixed language identification results from our experiments. Table 6 presents the macro and weighted score results for precision, recall, and F1 for each model on the test set. In addition, Table 7 illustrates the comparison of the average F1 scores by category for the proposed models.

We infer that the CRF model performs well in determining languages from code-mixed texts. The CRF model achieves macro and weighted F1 scores of 92.94% and 95.69%, respectively. Also, the CRF shows good performance in identifying individual languages for non-intra-word code-mixing. Compared to the CRF model, there is no significant improvement in the BLSTM-based architectures. Furthermore, adding a CRF layer on top of BLSTM cannot improve the F1 score. However, combining word and character CNN embedding can slightly enhance the macro F1 score. Among the BLSTM-based architectures, the BLSTM and BLSTM + character CNN models indicate comparable results with macro F1 scores of 91.06% and 91.11%, respectively. In addition, both models achieve the same value for weighted F1 of 93.66%.

Table 7 Category-wise comparison of F1 scores for each model.

Label	CRF	BLSTM	BLSTM-CRF	BLSTM + character LSTM	BLSTM + character CNN	mBERT	IndoBERTweet
EN	94.17	90.38	85.84	92.07	90.18	92.39	93.60
ID	97.11	95.45	95.05	95.94	95.58	96.50	97.06
JV	92.89	90.14	88.07	89.73	89.9	90.89	92.26
MIX_ID_EN	93.66	92	91.92	91.33	92.21	94.86	95.55
MIX_ID_JV	86.74	86.13	85.82	82.95	85.68	89.11	87.60
MIX_JV_EN	91.19	90.37	90.14	91.25	91.38	92.87	92.94
OTH	94.84	92.94	92.52	93.76	92.85	94.12	95.71

Our experiments with the fine-tuned BERT models indicate better results than the BLSTM-based models. The results provided by the fine-tuning BERT models demonstrate competitive performance compared to the other techniques. The fine-tuning IndoBERTweet model achieves the highest macro F1 score of 93.53% among the other models. This score is higher by 0.57% than the fine-tuning using mBERT, which gains a 92.96% F1 score. Further, the fine-tuning of BERT models proves an excellent achievement in identifying intra-word code-mixing.

Error prediction analysis

This sub-section discusses the error prediction analysis generated by the models. We randomly pick a sample from the test set to be analyzed as follows:

***Tweet*:** *malah gua yg diblock, barusan bgt tb2 profilenya gabisa dibuka wkwkwk*

***English*:** *instead, I was blocked, just now suddenly the profile can't be opened lol*

From the example above, the token '*profilenya*' is mixed Indonesian-English (MIX_ID_EN). It consists of '*profile*' (English) and the Indonesian suffix '*-nya*'. Among all models, the fine-tuned IndoBERTweet model can correctly identify such a token as mixed Indonesian-English. However, the remaining models identify it as an Indonesian word. This misclassification happens due to the spelling similarity between Indonesian and English. The term '*profile*' is written as '*profil*' in Indonesian, while in English, it is written as '*profile*'.

In addition, our models sometimes get confused when identifying between mixed Indonesian-English or Javanese-English. This confusion occurs since Indonesian and Javanese share similar morphemes. For example, the prefix '*di-*' in the following tokens, such as '*didownload*' (downloaded), '*dicancel*' (cancelled), '*diclick*' (clicked), and '*diprint*' (printed) can be identified as mixed Indonesian-English or mixed Javanese-English, depending on the sentence's context.

Based on our investigation, the fine-tuning BERT and BLSTM-based models can generate more accurate language identification than the CRF. Both fine-tuning BERT and BLSTM-based models can grasp the meaning of each word depending on the context. Also, we do not need to define the features manually for fine-tuning and BLSTM-based models. However, the fine-tuning BERT models can produce better identification than the

BLSTM-based models. Unlike directional models, which read a text from either left-to-right or right-to-left, the BERT models can read a text or a sequence of words simultaneously in any direction. Hence, the BERT model can understand the meaning of the given sequence based on the context to its right and left.

On the contrary, relying on CRF's handcrafted features is unreliable, especially in determining intra-word code-mixing. For example, the CRF model fails to identify words that have suffix similarity with Javanese, such as *'-re'*, *'-ke'*, and *'-ne'*. As a result, some words like *'atmosphere'*, *'Singapore'*, *'alike'*, and *'airplane'* are identified as mixed Indonesian-Javanese (MIX_ID_JV) by the CRF model. This is due to the fact that such words were not encountered during the CRF training.

Furthermore, we observe that tokens with JV labels are incorrectly detected as ID occasionally. This misclassification happens because of the shared vocabularies between Indonesian (ID) and Javanese (JV), such as *'aku'* (I), *'bareng'* (together), *'kuat'* (strong), and *'buku'* (book). Our models sometimes incorrectly predict the English (EN) tokens as Indonesian (ID). Misclassification from EN to ID mainly occurs when Indonesian words surround an English word. Words recognized in Indonesian and English sometimes confuse the model to predict the correct label, for instance, *'internet'*, *'admin'*, *'hotel'*, and *'got'*. Apart from that, some English words that are written in Indonesian or abbreviated styles also make our model confused to predict the correct label, like *'kambek'* (come back), *'mensen'* (mention), *'gaes'* (guys), *'fllw'* (follow), and *'pls'* (please). The actual label for such words is English, but those words are predicted as Indonesian.

CONCLUSION AND FUTURE WORK

In this article, we developed a code-mixed Indonesian-Javanese-English corpus for language identification (IJELID) using Twitter data. To standardize the annotation results, we created an annotation guideline for annotators. Also, we discussed some challenges encountered during corpus creation. To build the code-mixed language identification models, we explored various techniques, such as fine-tuning BERT, BLSTM-based, and CRF. Our experiments indicate that the fine-tuned IndoBERTweet models can effectively identify languages from the code-mixed tweets. Furthermore, using sub-word language representation in BERT models can produce a reliable model to identify languages for code-mixed texts. In addition, the BERT models can infer the context of each word from the given sequence better than the other techniques.

Even though the results of this research using BERT have demonstrated excellent performance, some spaces can still be improved from our study. In this study, we performed fine-tuning on pre-trained multilingual and monolingual models. Therefore, we intend to build a pre-trained model from code-mixed texts. We aim to investigate the impact of utilizing a pre-trained code-mixed model compared to the pre-trained monolingual and multilingual models.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Universitas Islam Indonesia for providing high-performance computing resources for conducting the research reported in this paper. We also appreciate the annotators who have annotated and validated the data. This work is supported by Universiti Brunei Darussalam (Grant no. UBD/RSCH/1.18/FICBF(a)/2023/007).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by Universiti Brunei Darussalam (Grant no. UBD/RSCH/1.18/FICBF (a)/2023/007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Universiti Brunei Darussalam: UBD/RSCH/1.18/FICBF(a)/2023/007.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Ahmad Fathan Hidayatullah conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Rosyzie Anna Apong conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Daphne T.C. Lai conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Atika Qazi conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at Zenodo: Hidayatullah, Ahmad Fathan. (2022). Code-mixed Indonesian-Javanese-English Twitter Dataset (Version v1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7567573.

REFERENCES

Adilazuarda MF, Cahyawijaya S, Winata GI, Fung P, Purwarianti A. 2022. IndoRobusta: towards robustness against diverse code-mixed indonesian local languages. In: *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*. 25–34.

Aji A, Winata GI, Koto F, Cahyawijaya S, Romadhony A, Mahendra R, Kurniawan K, Moeljadi D, Prasojo RE, Baldwin T. 2022. One Country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7226–7249.

Ansari MZ, Beg MS, Ahmad T, Khan MJ, Wasim G. 2021. Language identification of hindi-english tweets using code-mixed BERT. In: *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. Piscataway: IEEE, 248–252.

Ansari MA, Govilkar S. 2018. Sentiment analysis of mixed code for the transliterated Hindi and Marathi texts. *International Journal on Natural Language Computing* **7**:15–28 DOI 10.5121/ijnlc.2018.7202.

Arianto D, Budi I. 2020. Aspect-based sentiment analysis on indonesia's tourism destinations based on google maps user code-mixed reviews (study case: borobudur and prambanan temples). In: *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. 359–367.

Barik AM, Mahendra R, Adriani M. 2019. Normalization of Indonesian-english code-mixed twitter data. In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong Association for Computational Linguistics, 417–424.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**:37–46 DOI 10.1177/001316446002000104.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*. Minneapolis: Association for Computational Linguistics, 4171–4186.

Dowlagar S, Mamidi R. 2022. CMNEROne at SemEval-2022 Task 11: code-mixed named entity recognition by leveraging multilingual data. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 1556–1561.

Eberhard DM, Simons GF, Fennig CD. 2021. *Ethnologue: languages of the world*. Twenty-Fourth Edition. Dallas: SIL International.

Gambäck B, Das A. 2014. On measuring the complexity of code-mixing. In: *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*. 1–7.

Gundapu S, Mamidi R. 2018. Word level language identification in english telugu code mixed data. In: *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Hong Kong*.

Hidayatullah AF. 2015. Language tweet characteristics of Indonesian citizens. In: *2015 International Conference on Science and Technology (TICST)*. Piscataway: IEEE, 397–401.

Hidayatullah AF, Cahyaningtyas S, Pamungkas RD. 2020. Attention-based CNN-BiLSTM for dialect identification on Javanese text. In: *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. 317–324 DOI 10.22219/kinetik.v5i4.1121.

Hidayatullah AF, Qazi A, Lai DTC, Apong RA. 2022. A systematic review on language identification of code-mixed text: techniques, data availability,

challenges, and framework development. *IEEE Access* **10**:122812–122831 DOI 10.1109/ACCESS.2022.3223703.

Hoffmann C. 2014. *Introduction to bilingualism*. United Kingdom: Taylor & Francis.

Huang Z, Xu W, Yu K. 2015. Bidirectional LSTM-CRF models for sequence tagging. ArXiv preprint. arXiv:1508.01991.

Jamatia A, Das A, Gambäck B. 2018. Deep learning-based language identification in english-hindi-bengali code-mixed social media corpora. *Journal of Intelligent Systems* **28**:399–408 DOI 10.1515/jisys-2017-0440.

Jamatia A, Das A, Gambäck B. 2019. Deep learning-based language identification in english-hindi-bengali code-mixed social media corpora. *Journal of Intelligent Systems* **28**:399–408 DOI 10.1515/jisys-2017-0440.

Joshi R, Joshi R. 2020. Evaluating input representation for language identification in hindi-english code mixed text. ArXiv preprint. arXiv:2011.11263.

Kalita NJ, Saharia N. 2018. Language identification on code-mix social text. In: *Proceedings of the International Conference on Computing and Communication Systems*. Cham: Springer, 433–440.

Koto F, Lau JH, Baldwin T. 2021. IndoBERTweet: a pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics*. 10660–10668.

Kusampudi SSV, Chaluvadi A, Mamidi R. 2021. Corpus creation and language identification in low-resource code-mixed telugu-english text. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 744–752.

Lafferty J, McCallum A, Pereira FCN. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. 282–289.

Lamabam P, Chakma K. 2016. A language identification system for code-mixed English-Manipuri Social Media text. In: *2016 IEEE International Conference on Engineering and Technology (ICETECH)*. Piscataway: IEEE, 79–83.

Liu G, Guo J. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* **337**:325–338 DOI 10.1016/j.neucom.2019.01.078.

Mahata S, Das D, Bandyopadhyay S. 2021. Sentiment classification of code-mixed tweets using bi-directional RNN and language tags. In: *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. Kyiv*. 28–35.

Mahata SK, Mandal S, Das D, Bandyopadhyay S. 2019. Code-mixed to monolingual translation framework. In: *ACM International Conference Proceeding Series: Association for Computing Machinery*. 30–35.

Mandal S, Singh AK. 2018. Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture. In: *Proceedings of the 2018 EMNLP Workshop W-NUT: the 4th Workshop on Noisy User-generated Text*. Brussels: Association for Computational Linguistics, 116–120.

Mave D, Maharjan S, Solorio T. 2018. Language identification and analysis of code-switched social media text. In: *Proceedings of the third workshop on computational approaches to linguistic code-switching*. Melbourne: Association for Computational Linguistics, 51–61.

Nguyen L, Bryant C, Kidwai S, Biberauer T. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data* **7**:7 DOI 10.5334/johd.44.

Phadte A, Wagh R. 2017. Word level language identification system for Konkani-english code-mixed social media text (CMST). In: *Proceedings of the 10th Annual ACM India Compute Conference*. Bhopal: Association for Computing Machinery, 103–107.

Poostchi H, Borzeshi EZ, Piccardi M. 2018. In: *BiLSTM-CRF for persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated persian dataset. LREC*. Miyazaki: European Language Resources Association (ELRA).

Reimers N, Gurevych I. 2017. Optimal Hyperparameters for deep LSTM-networks for sequence labeling tasks. ArXiv preprint. arXiv:1707.06799.

Ritchie WC, Bhatia TK. 2012. Social and psychological factors in language mixing. In: *The handbook of bilingualism and multilingualism*. United Kingdom: Wiley, 375–390.

Rizal AN, Stymne S. 2020. Evaluating word embeddings for Indonesian–English code-mixed text based on synthetic data. In: *Proceedings of the LREC 2020—4th Workshop on Computational Approaches to Code Switching. Marseille, France*. 26–35.

Sabty C, Mohamed I, Ö Çetinoğlu, Abdennadher S. 2021. Language identification of intra-word code-switching for Arabic-English. *Array* **12**:100104 DOI 10.1016/j.array.2021.100104.

Samih Y, Maharjan S, Attia M, Kallmeyer L, Solorio T. 2016. Multilingual code-switching identification *via* LSTM recurrent neural networks. In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin: Association for Computational Linguistics, 50–59.

Sarma N, Singh RS, Goswami D. 2022. SwitchNet: learning to switch for word-level language identification in code-mixed social media text. *Natural Language Engineering* **28**:337–359 DOI 10.1017/s1351324921000115.

Shashirekha H, Balouchzahi F, Anusha M, Sidorov G. 2022. CoLI-machine learning approaches for code-mixed language identification at the word level in Kannada-English texts. ArXiv preprint. arXiv:2211.09847.

Shekhar S, Sharma DK, Beg MMS. 2020. An effective Bi-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and Roman Hindi. *Computacion y Sistemas* **24**:1415–1427 DOI 10.13053/cys-24-4-3151.

Singh K, Sen I, Kumaraguru P. 2018. Language identification and named entity recognition in hinglish code mixed tweets. In: *Proceedings of ACL 2018, Student Research Workshop*. Melbourne: Association for Computational Linguistics, 52–58.

Smith I, Thayasivam U. 2019. Language detection in sinhala-english code-mixed data. In: *2019 International Conference on Asian Language Processing (IALP)*. Piscataway: IEEE, 228–233.

Suciati A, Budi I. 2019. Aspect-based opinion mining for code-mixed restaurant reviews in Indonesia. In: *2019 International Conference on Asian Language Processing (IALP)*. Piscataway: IEEE, 59–64.

Sutton C, McCallum A. 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning* **4**:267–373 DOI 10.1561/2200000013.

Thara S, Poornachandran P. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access* **9**:118837–118850 DOI 10.1109/access.2021.3104106.

Tho C, Heryadi Y, Lukas L, Wibowo A. 2021. Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach. In: *2nd Annual Conference of Science and Technology (ANCOSET 2020)*. Malang: IOP Publishing Ltd,.

Tonja AL, Yigezu MG, Kolesnikova O, Tash MS, Sidorov G, Gelbuk A. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. ArXiv preprint. arXiv:2211.14459.

Ulčar M, Robnik-Šikonja M. 2020. Finest bert and crosloengual bert. In: *International Conference on Text, Speech, and Dialogue*. Cham: Springer, 104–111.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Ł Kaiser, Polosukhin I. 2017. Attention is all you need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach: 5998–6008.

Winata GI, Aji AF, Cahyawijaya S, Mahendra R, Koto F, Romadhony A, Kurniawan K, Moeljadi D, Prasojo RE, Fung P. 2022. NusaX: multilingual parallel sentiment dataset for 10 Indonesian local languages. ArXiv preprint. arXiv:2205.15960.

Wintaka DC, Bijaksana MA, Asror I. 2019. Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF. *Procedia Computer Science* **157**:221–228 DOI 10.1016/j.procs.2019.08.161.

Wolf T, Chaumond J, Debut L, Sanh V, Delangue C, Moi A, Cistac P, Funtowicz M, Davison J, Shleifer S. 2020. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics*. 38–45.

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K. 2016. Google's neural machine translation system: bridging the gap between human and machine translation. ArXiv preprint. arXiv:1609.08144.

Yirmibeşoğlu Z, Eryiğit G. 2018. Detecting code-switching between turkish-english language pair. In: *Proceedings of the 2018 EMNLP Workshop W-NUT: the 4th workshop on noisy user-generated text*. Brussels: Association for Computational Linguistics, 110–115.

Yulianti E, Kurnia A, Adriani M, Duto YS. 2021. Normalisation of Indonesian-English code-mixed text and its effect on emotion classification. *International Journal of Advanced Computer Science and Applications* **12**:674–685.

Zhang X, Zhao J, LeCun Y. 2015. Character-level convolutional networks for text classification. 649–657ArXiv preprint. arXiv:1509.01626.

PeerJ Computer Science

BERT-based tourism named entity recognition: making use of social media for travel recommendations

Dhomas Hatta Fudholi, Annisa Zahra, Septia Rani, Sheila Nurul Huda, Irving Vitra Paputungan and Zainudin Zukhri

Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

ABSTRACT

Background. Social media has become a massive encyclopedia of almost anything due to its content richness. People tell stories, write comments and feedback, and share knowledge through social media. The information available on social media enables 'clueless' travelers to get quick travel recommendations in the tourism sector. Through a simple query, such as typing 'places to visit in Bali', travelers can get many blog articles to help them decide which places of interest to visit. However, doing this reading task without a helper can be overwhelming.

Methods. To overcome this problem, we developed Bidirectional Encoder Representations from Transformers (BERT)-based tourism named entity recognition system, which is used to highlight tourist destination places in the query result. BERT is a state-of-the-art machine learning framework for natural language processing that can give a decent performance in various settings and cases. Our developed tourism named entity recognition (NER) model specifies three different tourist destinations: heritage, natural, and purposefully built (man-made or artificial). The dataset is taken from various tourism-related community articles and posts.

Results. The model achieved an average F1-score of 0.80 and has been implemented into a traveling destination recommendation system. By using this system, travelers can get quick recommendations based on the popularity of places visited in the query frame.

Discussion. Based on the survey that we conducted to target respondents who have never visited and have no or limited knowledge about tourist attractions in some example cities, their average interest level from the recommendation results is higher than four on a scale of 1 to 5. Thus, it can be considered a good recommendation. Furthermore, the NER model performance is comparable to another related research.

Subjects Artificial Intelligence, Data Mining and Machine Learning, Data Science, Natural Language and Speech, Text Mining

Keywords BERT, Named entity recognition, Traveling, Recommendation

Submitted 10 June 2022
Accepted 9 November 2023
Published 21 December 2023

Corresponding author
Dhomas Hatta Fudholi,
hatta.fudholi@uii.ac.id

Academic editor
Kaize Shi

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj-cs.1731**

Copyright
2023 Fudholi et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

For some people, traveling is a thrilling experience. It allows them to try new things like visiting new places, meeting new people, and learning about the area's culture. According to World Tourism Organization (UNWTO) figures, the number of international tourists

arriving in the United States grew from 2017 to 2019 (*UNWTO, 2020*). Due to the COVID-19 epidemic, it declined substantially in 2020. On the other hand, global tourism increased by 4% in 2021 compared to the previous year. International visitor visits could increase by 30% to 78% in 2022 compared to 2021 and only return to 2019 levels in 2024 or later, according to UNWTO expert forecasts (*UNWTO, 2022*).

People usually plan which destinations they wish to see before going on vacation. Various Internet or social media sources serve as their primary information and inspiration while planning a vacation. Social media has altered the way people interact and connect, including travel, during the last decade. Friends, travel bloggers, and influencers use Instagram, Facebook, Twitter, and other social media to share their vacation photographs and tales. According to a survey, 87 percent of millennials used Facebook to plan their vacations, and more than half utilized Pinterest or Twitter (*Condor Ferries, 2023*). According to another study conducted by Schofields, over 40% of millennials aged 18–33 consider "Instagrammability" while choosing a vacation destination (*Schofield, 2017)*). Such social media messages have a definite impact on a traveler's decision to go where they want to go next. However, making a decision can be difficult and time-consuming, especially when visiting new regions or traveling alone. Furthermore, the abundance of sources frequently leads to new issues. It is understandable that carefully analyzing all sources to gather useful information can be overwhelming.

Named entity recognition (NER) is a method that is believed to solve these issues. NER will extract and classify information from the text available on any Internet source or social media platform based on specified tourism entities or categories to deliver attraction recommendations. NER assigns labels (classes) or semantic categories to words or phrases found in a text (*Mansouri, Affendey & Mamat, 2008*). This type of recognition helps determine what users look for when they search (*Eftimov, Seljak & Korošec, 2017*).

In NER task, there are three basic approaches: lexicon-based, rule-based, and machine learning approaches. Maximum entropy, support vector machines (SVM), hidden Markov model (HMM), conditional random fields (CRF), decision tree, and the hybrid method are the examples of these approaches (*Kaur & Gupta, 2020*). Besides CRF that is well-known to be the model in NER task (*Agrawal et al., 2022*), recurrent neural network (RNN), including its successor, long short-term memory (LSTM), becomes the pioneer of DL architecture that handles sequence modeling, as in sentences. However, a new approach that can be used in NER modeling task, bidirectional encoder representations from transformers (BERT). BERT is a self-attention-based language model representation pre-trained on several language modeling tasks utilizing raw unlabeled texts (*Taher, Hoseini & Shamsfard, 2020*). BERT can blend the left and right context that enables the pretraining of a bidirectional deep transformer. It removes unidirectional limitation. The fine-tuned, pre-trained BERT models might perform well without external resources or feature extraction, and it is provenly found to be effective (*Agrawal et al., 2022*). In *Agrawal et al. (2022)*, BERT-based model outperformed other models such as CRF and the combination of bidirectional long short-term memory (Bi-LSTM) with the CRF model.

NER has been researched in a variety of domains. The BiLSTM-CRF model is utilized in the biomedical arena to recognize drug names in biomedical publications (*Zeng et al., 2017*). The input vector for this investigation was created by concatenating word embedding and character-level embedding. In the biological area, (*Boudjellal et al., 2021*) used an Arabic dataset to do NER and found that employing a pre-trained monolingual BERT case model rather than a multilingual BERT case model yielded the best results. *Sobhana, Mitra & Ghosh (2010)* uses the IITKGP-GEOCORP dataset with the CRF approach to work on NER in the geological domain. The IITKGP-GEOCORP is a geology-related data set compiled from various publications and journals published in India. A NER model was created in the archeology area to annotate text collections with archeological named entities (*Brandsen, Verberne & Lambers, 2022*). An approximately 10 GB of Mongolian dataset used in *Cheng et al. (2020)* contained various Mongolian news and Mongolian Tourism obtained from various websites. *Bouabdallaoui et al. (2021)* built a Moroccan dataset that was collected from TripAdvisor Moroccan forum which contained 46,120 topic threads. The Korean and English tourism datasets in *Jwa & Jwa (2022)* were constructed using tourism information data from the VisitJeju web of the Jeju Tourism Organization, internet websites, and smart tourism information system. Their dataset contains 20,768 words for Korean NER learning data, while the English NER learning data has 12,024 words. NER research has also been done in a variety of languages, including Portuguese (*Souza, Nogueira & Lotufo, 2019*), Turkish (*Güneş & Tantuğ, 2018*), and Bengali (*Drovo et al., 2019*). In *Labusch, Neudecker & Zellhöfer (2022)*, German datasets were used. They used BERT to create the model. Their findings show that a properly trained BERT model may perform well in various of contexts and even deliver state-of-the-art performance in many circumstances without requiring substantial fine-tuning and optimization. *Hakala & Pyysalo (2019)* performed NER on the Spanish dataset. They used a CRF-based baseline technique and multilingual BERT to complete the PharmaCoNER task on Spanish biological named entity recognition. For NER in Persian, *Taher, Hoseini & Shamsfard (2020)* employed the BERT model. For identifying named entities in the Chinese language, *Zhuang et al. (2021)* used a bidirectional gated recurrent unit with CRF. To produce high-quality word vectors, they also use a pre-trained BERT model. *Gao, Zheng & Zhao (2021)* combined BERT with BiLSTM and CRF and used a Chinese dataset. The BERT-BiLSTM-CRF model outperforms the Word2vec-BiLSTM-CRF model, according to the results. Their best model can correctly recognize various entity categories.

NER has been used in various works in the tourism domain (*Cheng et al., 2020*; *Vijayakrishna & Sobha, 2008*; *Saputro, Kusumawardani & Fauziati, 2016*; *Hu, Nuo & Tang, 2019*; *Chantrapornchai & Tunsakul, 2021*; *Xue et al., 2019*). CRF was used in *Vijayakrishna & Sobha (2008)* to perform tourism named entity recognition using Tamil dataset which obtained an F1-score of 80.44%. A semi-supervised NER was developed in *Saputro, Kusumawardani & Fauziati (2016)* to find new tourist attractions. However, another two-stage idea (YATSI), naive Bayes classifier (NBC), and K-nearest neighbor (KNN) are the methodologies used in that research and it obtained an average F1-score of 69%. Using BERT and other techniques, (*Hu, Nuo & Tang, 2019*) extracts the Chinese Tourism field attribute, such as location, time, and Chinese name; *Chantrapornchai & Tunsakul*

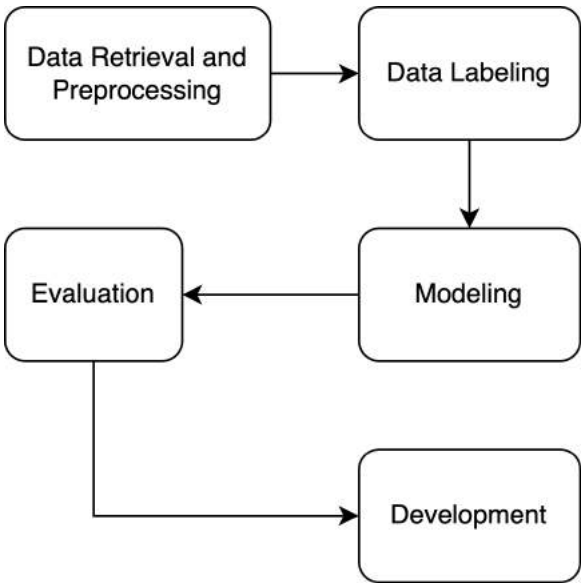


Figure 1 **The research methodology of the study.**
Full-size DOI: 10.7717/peerjcs.1731/fig-1

(2021) extracts location, accommodation, and facility from TripAdvisor, Traveloka, and Hotels.com; *Xue et al. (2019)* extracts person, location, organization, and time from Trip.com and Mafengwo app. Recently, BERT has been used in *Cheng et al. (2020)* to create a corpus for Mongolian tourism and obtained more than 80% score. BERT looks promising for NER.

A BERT-based NER model is proposed in this article to support tourism destination recommendations from social media posts. As there is still no similar research that we found, this research becomes one of the pioneers in using BERT-based NER to be the core of recommendation engine. This research is also conducted to see BERT's performance for NER modeling on only English tourism dataset. In the Materials & Methods section, the materials and methods used in the study are described. Results and discussion are presented in the Results and Discussion. Finally, the Conclusions section concludes the paper.

MATERIALS & METHODS

The aim of our resaerch is to develop a recommendation engine in tourism domain that is based on BERT-based NER. Figure 1 shows the research flow. The research starts with data retrieval and preprocessing step. The required and related datasets are collected and is preprocessed to produce clean dataset. A clean dataset is then labeled with tourism categories. In the modeling step, BERT based model architecture is used. The developed model will be evaluated to see its performance. Finally, the best model build will be implemented into a proof-of-concept application.

Data retrieval and preprocessing

We gathered the data for this study by crawling the web for numerous English travel articles, such as worldofwanderlust.com, touropia.com, tourmyindia.com, vaticantour.com, myguidezimbabwe.com, daytours4u.com, and Traveloka.com. The search was initially conducted by entering several keywords alternately into the Google search engine, such as "top cities for tourists", "best places to visit", "top world heritage sites", "world heritage list", "best destinations for nature lovers", and "best natural tourist attractions". The data obtained up to that point was preprocessed, which included the removal of emoticons and URLs, as well as tokenization. We opted to remove the emoticons and URLs because they did not contribute anything to the process of recognizing a named entity. On the other hand, tokenization is breaking down sentences into their constituent elements, known as tokens. Tokenization is done at the word token level, where the analysis and inference process occur. We collected 183,507 tokens from 92 articles with 8,500 sentences and 17,137 different words.

Data labeling

Swarbrooke & Page (2002) proposed tourist attraction categories, namely natural, man-made/artificial purpose-built and non-tourist purpose-built, and special events. Included in the first category are tourist attractions in the form of natural products such as beaches, mountains, and forests. The second category includes tourist attractions explicitly built to attract tourists, such as parks, museums, and galleries. The third category contains tourist attractions whose development goals are not to attract tourists, for example, historical buildings or places such as monuments, temples, and castles. Cathedrals and churches, which are places of worship, also fall into the third category. The last category is festivals and other events, such as sporting events.

Such labels are partially adopted in our study. Natural is used to classify places with rivers, waterfalls, mountains, beaches, and other places that offer natural scenery that is open to the public. Man-made purpose-built and man-made non-tourist purpose-built categories are combined into a single category, called Purpose. It covers purposefully built (man-made) tourist attractions, such as amusement parks and museums. An additional category is Heritage. Tourist attractions that are classified as Heritage are those that have been around for many years. They are historical, cultural, ancient, or representatives of a culture, such as castles, mosques, ruins, and cathedrals. In contrast to *Swarbrooke (2002)*, we do not specifically have a Special event category due to the nature of such an event that is not always available.

We employed the IOB tagging format in this work, with B-label indicating a token at the start of a noun phrase, an I-label indicating a token within the current noun phrase, and O-label indicating otherwise. B-prefix correlates to the first word of a tourist attraction's name in this study, whereas I-prefix corresponds to the second through the last word of a tourist attraction's name. The labeling process is done manually by the author. The total number of tokens for each label is shown in Table 1. We have reasonably balanced entities in each of the three tourist attraction categories. Due to the nature of expressing descriptive

Table 1 Number of tokens.

Label	Number of tokens
O	171,728
B-NATURAL	1,789
I-NATURAL	1,853
B-HERITAGE	1,401
I-HERITAGE	2,051
B-PURPOSE	1,811
I-PURPOSE	2,874

articles, the mentioned tourist attraction entity in the dataset has a small portion compared to other non-tourist attraction tokens.

Modeling

In this study, we used a model that had been pre-trained utilizing BERT to perform the NER task. It was introduced in *Devlin et al. (2018)* as a technique for pre-training deep bidirectional representations from the unlabeled text. The pre-training was accomplished by simultaneously conditioning all model layers on the right and left contexts.

As BERT depends on the tokenization at the word piece level, we decided to initially perform tokenization at that level before putting the data into the previous trained version of the BERT model. For instance, the word 'Prehistoric', which has the label 'I-PURPOSE', is tokenized to 'Pre', '##his', '##tor', and '##ic'. The label for each of these word pieces will be 'I-PURPOSE'. After the tokenization process is done to all sentences, we add the specialized BERT tokens ('CLS' and 'SEP'). Then, based on the maximum length, we padded or shortened the tokens, and for the next phase, we built the attention mask in addition to the labels using the dictionary that had been defined before this stage. A training dataset and a testing dataset were created from the original dataset used in this work. We adopted an 80:20 split, with 80 percent of the data going toward training and 20 percent testing. We used holdout method for the splitting mechanism.

To select the best model, we applied hyperparameter tuning to the model architecture during the modeling process. Different learning rates, epochs, and batch sizes were tested. We also used two distinct BERT models, BERT-base-cased and BERT-base-uncased, to test our hypothesis. We modify one type of hyperparameter at a time during the experiment while the other hyperparameters are frozen. The base hyperparameter that we used is as follows: learning rate = 0.00001, epoch = 3, training batch size = 4, testing batch size = 2, model = BERT-base-cased. On top of that, Adam optimizer is used. The hyperparameter optimization performed on the testing set.

Evaluation

After completing the training procedure, we used the testing dataset to evaluate our model's performance. The F1-score is used to assess its performance. Because most tokens in the NER dataset are labeled as O, which implies that the token is not a named entity, F1-score is a better choice than accuracy (*Seok et al., 2016*). The F1-score is the harmonic mean of

precision and recall, as shown by Eq. (1). The F1-score ranges from 1 to 0, such that 1 being the highest and 0 being the lowest.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}. \quad (1)$$

In addition to the F1-score, a survey on potential users is conducted as part of the evaluation process. The target audience for this survey is 'clueless' visitors or those who have little or no information about the city and the tourist attractions they plan to visit. This survey aims to see if the system's recommendation meets the user's needs, in other words, how appealing the tourist attractions recommendation is to the user. A 1 to 5 rating scale is used in the survey to capture user interest in the suggested outcome. A rating of 1 indicates that the recommendation is not interesting, while a rating of 5 indicates that the recommendation is very interesting. In the questionnaire, we choose four cities as an example: Kyoto, London, Lisbon, and Colorado. The average user interests for each city is then used to determine the survey results.

Development

As a proof of concept in delivering the recommendations, a web-based application is built using the best model. The application accepts user query that asks for the tourist destination location that has never been visited. The program will list tourist places in the specified three destination categories based on information acquired from community articles on the web.

RESULTS

Finding the best hyperparameter setup that results in the best model in terms of F1-score is the first step in the modeling process. Table 2 shows the final hyperparameter using BERT-based-case weightage. The BERT-base-cased model has a linear layer on top of the output of the hidden states, and this layer functions as a token classification layer. These hyperparameters resulting the best model with average F1-score of 0.80. Hyperparameters in Table 2 are chosen through experimenting with different values of each type of hyperparameter. We tried 0.0001 and 0.000001 as the other values in terms of learning rates. By doing so, the F1-score is degraded to 0.61 and 0.29, respectively. Raising the epoch from 2 to 5 does not give significant differences. Hence, we picked 3 as the value of the epoch number. In the batch size sections, we tried to double up the values of the training batch size to 8, 16, and 32, and the testing batch size to 4, 8, and 16. The F1-score result degrades to 0.78, 0.75, and 0.66. Finally, using the BERT-base-uncased model also slightly reduced the F1-score to 0.78.

Table 3 shows the performance metrics score for all categories, including precision, recall, and F1-score. Heritage and Natural have a slightly better F1-score than Purpose because they tend to have similar patterns, such as the words that are often used in naming tourist attractions, for example both "Lake Tahoe" and "Lake Toba" used the word "Lake" in their names. So that the model can more easily detect these tourist attractions. Meanwhile, there are more word variations for tourist attraction that fall into the Purpose

Table 2 Hyperparameter used.

Hyperparameter	Value
Training batch size	4
Testing batch size	2
Epoch	3
Optimizer	Adam
Learning rate	0.00001

Table 3 Performance metrics score for each tourist attraction categories.

	Precision	Recall	F1-score	Support
HERITAGE	0.81	0.81	0.81	521
NATURAL	0.78	0.85	0.81	796
PURPOSE	0.74	0.80	0.77	722

Table 4 Example of prediction result for new sentence.

Token	Predicted label
Grand	B-NATURAL
Lake	I-NATURAL
is	O
just	O
before	O
the	O
western	O
entrance	O
to	O
Rocky	B-PURPOSE
Mountain	I-PURPOSE
National	I-PURPOSE
Park	I-PURPOSE
.	O

category. The variances could confound the model and prevent it from making accurate predictions. We also tried to predict new named entities in new input sentences based on the trained model. The prediction result is shown in Table 4. We can see from the table that our model managed to predict all tokens correctly.

We also built a simple web-based application to implement the NER model for extracting tourist attractions from several articles. This web interface is useful for people to use the model that has been built so that they can get tourist recommendations in cities that are not well recognized by users (this is what we define as clueless travelers). As shown in Fig. 2,

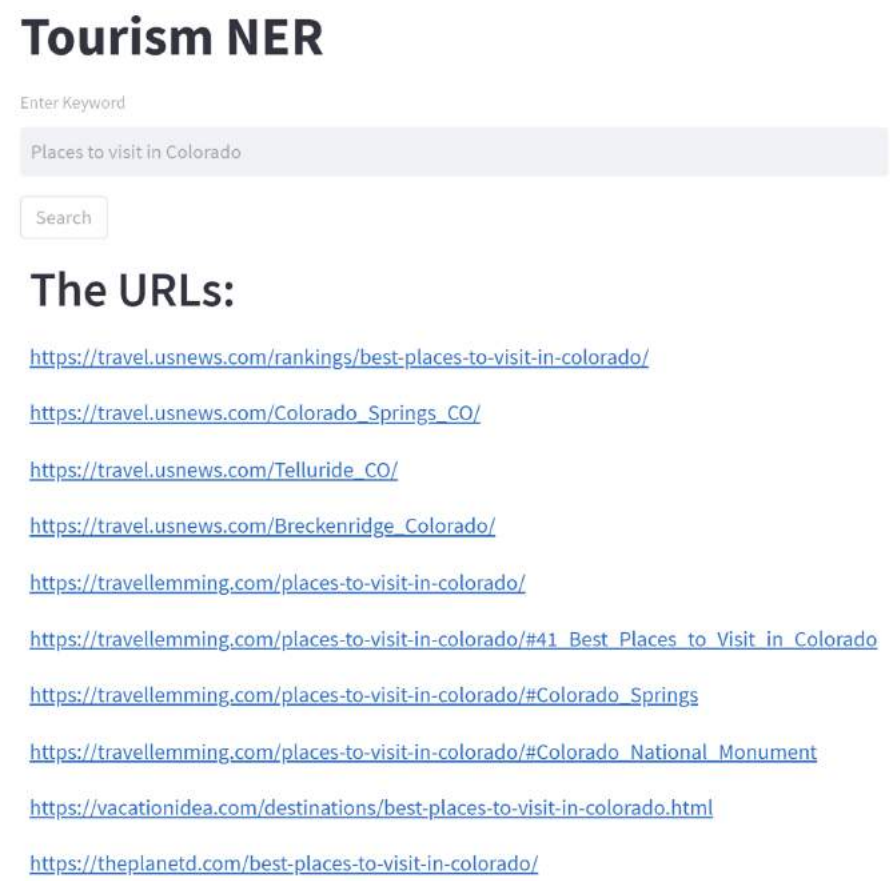
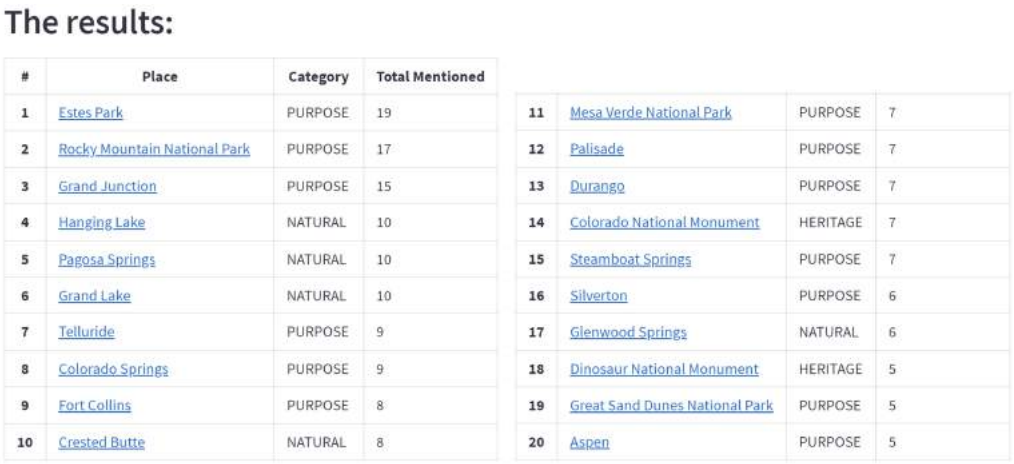


Figure 2 **The recommendation application with query sample and result.**
Full-size DOI: 10.7717/peerjcs.1731/fig-2

we tried "Places to visit in Colorado" as the input keyword. The URLs found, which will then be scraped, is also shown in Fig. 2.

We have created a web-based application to use the NER model to extract tourist attractions from various of articles. This web interface allows people to utilize the model created to acquire tourist recommendations in cities that are not well known by users (this is what we define as clueless travelers). We attempted "Places to visit in Colorado" as an input keyword, as shown in Fig. 2. Figure 2 also shows the URLs that hold tourist destination information to be collected later.

The contents of the articles in the URLs were scraped automatically after the URLs were retrieved from Google Search Engine. We can use this method to search travel blogs and reviews as sources. Each word with the entity labels Heritage, Natural, or Purpose will be recorded. The number of occurrences in the article sources is then used to rank the terms. The tourist attractions located in the article sources are presented to the user as



The results:

#	Place	Category	Total Mentioned
1	Estes Park	PURPOSE	19
2	Rocky Mountain National Park	PURPOSE	17
3	Grand Junction	PURPOSE	15
4	Hanging Lake	NATURAL	10
5	Pagosa Springs	NATURAL	10
6	Grand Lake	NATURAL	10
7	Telluride	PURPOSE	9
8	Colorado Springs	PURPOSE	9
9	Fort Collins	PURPOSE	8
10	Crested Butte	NATURAL	8
11	Mesa Verde National Park	PURPOSE	7
12	Palisade	PURPOSE	7
13	Durango	PURPOSE	7
14	Colorado National Monument	HERITAGE	7
15	Steamboat Springs	PURPOSE	7
16	Silverton	PURPOSE	6
17	Glenwood Springs	NATURAL	6
18	Dinosaur National Monument	HERITAGE	5
19	Great Sand Dunes National Park	PURPOSE	5
20	Aspen	PURPOSE	5

Figure 3 **The example of detected tourist attraction in the recommendation application.**
Full-size DOI: 10.7717/peerjcs.1731/fig-3

Table 5 **Evaluation result from survey.**

City	Average of users' interest
Kyoto	4.18
London	4.01
Colorado	4.14
Lisbon	4.08

the suggestion output after being ranked. The recommendation results for the keyword "Places to visit in Colorado" are given in Fig. 3.

We are not considering whether the review is good or negative in this study. On the other hand, given the keywords selected, good review outcomes are more likely than negative review results. In order to evaluate the suggested results, we ran a survey utilizing a questionnaire with target respondents who had never been and had little or limited information about tourist sites in specific example locations, such as London, Kyoto, Colorado, and Lisbon. These criteria are set to portray the situation and behavior of clueless travelers.

The survey was completed by 40 people ranging in age from 17 to 54. They never went to all cities listed as examples: London, Kyoto, Colorado, and Lisbon. On a scale of 1 to 5, the respondent judged how good is the recommendation result. Table 5 shows the average level of user interest in the sample cities' recommended tourist attractions. With the average interest level higher than 4 on a scale of 1 to 5, the output from NER in those source articles can be considered a good recommendation.

DISCUSSION

As a technique to label a phrase inside a sentence, NER can be utilized in this research to give a tourist attraction recommendation. The recommendation gathered from this

approach is a general recommendation that has not considered the user's personality and preferences. Thus, this kind of recommendation is still suitable for initial ideas of the most attractive places, especially for travelers with no or limited initial knowledge of the city they want to visit. Users do not have to search and read many articles and blogs to gather a list of tourist attraction recommendations.

We have developed a NER model in tourism to detect tourism destinations from any English articles and implemented the model as recommendation engine. The recommendation result is proven to be a good recommendation from the survey conducted with the average interest level higher than 4 in a scale of 1 to 5. However, our approach has some limitations. This study has not yet considered whether the article sources have positive or negative reviews of the tourist destinations mentioned in them. On the other hand, this can be further exploited as an advantage by using positive or negative keywords to gather the desired results. Another limitation is that some article sources may write similar tourist location names in different writing. For example, when using the keyword 'Places to visit in Lisbon', the system gave several recommendations, two of them are Jerónimos Monastery and Jeronimos Monastery. Both are treated as two different recommendations, even though the two phrases refer to the same tourist destination. The situation happens because both ways of writing (Jeronimos Monastery/Jerónimos Monastery) are commonly used in the article sources. Our further research will try to detect the similarity in the tourist destination names to avoid such redundancy.

Despite these limitations, the recommendations from our approach are pretty suitable for clueless tourists as our target audience. Our previous research (*Rani et al., 2021*) only targeted specific users by considering their personalities and preferences. The results from this research complement our previous research to give more elaborate recommendations for tourists. Furthermore, comparing our current BERT-based NER model to our previous study and experiments in developing NER model using BiLSTM-CRF (*Zahra, Hidayatullah & Rani, 2022*), our BERT-based model is proven to deliver better performance (80% in F1-score) rather than the BiLSTM-CRF-based model (75.25% in F1-score).

If we focus on the NER model itself, study (*Cheng et al., 2020*; *Bouabdallaoui et al., 2021*; *Jwa & Jwa, 2022*; *Chantrapornchai & Tunsakul, 2021*) has the same purpose as ours in developing the NER model for the tourism domain using BERT-based architecture. Table 6 shows the previous studies and highlights four key points of the studies: datasets, entities, methods, and performance. In the view of datasets, most studies use tourism-related websites and applications to get the data, including ours. While our study is more global in terms of the scope of tourism datasets, other studies specifically focus on tourism in: Mongolian (*Cheng et al., 2020*), Moroccan (*Bouabdallaoui et al., 2021*), Korea (*Jwa & Jwa, 2022*), and Thailand (*Chantrapornchai & Tunsakul, 2021*). The number of the dataset cannot be compared directly to each other, studies (*Cheng et al., 2020*) uses the number of sentences, (*Bouabdallaoui et al., 2021*) uses the number of topic threads, (*Jwa & Jwa, 2022*) and (*Chantrapornchai & Tunsakul, 2021*) use the number words/keywords. Our dataset has 8,500 sentences which is half the number of sentences compared to the dataset in *Cheng et al. (2020)*. The target entities to be extracted by the developed NER model are mostly tourism-related destinations or places, even though the label might be different.

Table 6 Previous studies of BERT-based NER in tourism domain.

Study	Dataset	Entities	Method	Performance (F1-Score)
Cheng et al. (2020)	Mongolian news and tourism websites (16,000 sentences)	Person, scenic, cultural, organization, spesific field	CRF, BiLSTM-CRF, BERT	78.55%, 80.08%, 82.09%
Bouabdallaoui et al. (2021)	Moroccan TripAdvisor forum (46,120 topic thread)	Regions, hotels, restaurants, shopping places, gastronomy, religion, outfits, transports and touristic guides	BERT, RoBERTa, XML-RoBERTa	70.5%, 69.44%, 68.53%
Jwa & Jwa (2022)	Smart tourism app, visitJeju web, and web search (20,768 words in Korean and 12,024 words in English)	Tourist destination, accomodation, restaurants and cafes	KoBERT-CRF	94%
Chantrapornchai & Tunsakul (2021)	Two different dataset for 8 different provinces in Thailand: 1. Tripadvisor, Traveloka, Hotels.com (58,828 keywords) 2. Tripadvisor website (29,619 keywords after cleaning)	Two different NER tasks: 1. Location, organization, and facility 2. Restaurant, hotels, shopping and tourism	BERT	1. 38%–91% (default tokenization) 2. 77,9% - 98,8%

Study (*Cheng et al., 2020*) has scenic and cultural categories, (*Bouabdallaoui et al., 2021*) has gastronomy and religion category, and our study has specific heritage, natural, and purposes categories. BERT has become the base architecture model in the studies, including ours. However, study (*Bouabdallaoui et al., 2021*) added more BERT-based models such as RoBERTa, and (*Jwa & Jwa, 2022*) uses KoBERT. Finally, in terms of performance, the F1-score results on all the studies vary. We cannot compare it apple to apple since the dataset and the category in are different. However, we can see that the BERT-based model for tourism NER studies averagely has F1-score ranging from about 70 to 90 percent, where our NER model has 80% as the score.

CONCLUSIONS

Social media and community information on the web in the tourism industry allows clueless travelers to acquire instant vacation recommendations. Travelers can find various blog entries that might assist them in picking which locations of interest to visit during their vacation by typing a basic query and getting automatic recommendations rather than manually DOIng reading tasks, which can be time-consuming.

We created BERT-based tourism named entity recognition model that highlights tourist destinations in one query search results to address this challenge. Our tourism NER model distinguishes three types of tourist destinations: heritage, natural, and purposely constructed (man-made). The data was compiled from various community

articles and postings about tourism. We picked the best model from our experiments through the hyperparameter tuning phase. The model has been integrated into a travel location recommendation system with an average F1-score of 0.80. Travelers can get rapid recommendations using this approach based on the popularity of destinations visited in the query context.

To evaluate the recommendation system, we ran a survey to target respondents who have never been and have little or limited information about tourist sites in selected sample cities. The average interest level from the suggestion results is higher than 4 on a scale of 1 to 5. As a result, it is a good recommendation. In addition, the NER model's performance is equivalent to that of other related studies. Considering future development of the recommendation system, we will use the model and the recommendation result to form a systematic itinerary recommendation for the user.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Dhomas Hatta Fudholi conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Annisa Zahra performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Septia Rani conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Sheila Nurul Huda conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Irving Vitra Paputungan conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Zainudin Zukhri conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The modeling code, implementation code, and data are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1731#supplemental-information.

REFERENCES

Agrawal A, Tripathi S, Vardhan M, Sihag V, Choudhary G, Dragoni N. 2022. BERT-based transfer-learning approach for nested named-entity recognition using joint labeling. *Applied Sciences* **12(3)**:976 DOI 10.3390/APP12030976.

Bouabdallaoui I, Guerouate F, Bouhaddour S, Saadi C, Sbihi M. 2021. Named entity recognition applied on Moroccan tourism corpus. *Procedia Computer Science* **198**:373–378 DOI 10.1016/J.PROCS.2021.12.256.

Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, Dai L. 2021. ABioNER: a BERT-based model for arabic biomedical named-entity recognition. *Complexity* **2021**:1–6 DOI 10.1155/2021/6633213.

Brandsen A, Verberne S, Lambers K. 2022. Can BERT dig it?—named entity recognition for information retrieval in the archaeology domain. *Journal on Computing and Cultural Heritage* **15**:1–18 DOI 10.1145/3497842.

Chantrapornchai C, Tunsakul A. 2021. Information extraction tasks based on BERT and SpaCy on tourism domain. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)* **15(1)**:108–122 DOI 10.37936/ecti-cit.2021151.228621.

Cheng X, Wang W, Bao F, Gao G. 2020. MTNER: a corpus for Mongolian tourism named entity recognition. In: *Machine Translation: 16th China Conference, CCMT 2020, Hohhot, China, October 10–12, 2020, Revised Selected Papers 16 (pp. 11–23)*. Springer Singapore.

Condor Ferries. 2023. Millennials Travel, Statistics & Trends 2023. *Available at https://www.condorferries.co.uk/millennials-travel-statistics-trends*.

Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019-2019 Conference of the North American Chapter Association of Computer Linguistics and Human Language Technology, vol. 1*. 4171–4186 DOI 10.48550/arxiv.1810.04805.

Drovo MD, Chowdhury M, Uday SI, Das AK. 2019. Named entity recognition in Bengali text using merged hidden Markov model and rule base approach. In: *2019 7th International Conference on Smart Computer Communication. ICSCC 2019* DOI 10.1109/ICSCC.2019.8843661.

Eftimov T, Seljak BK, Korošec P. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLOS ONE* **12(6)**:e0179488 DOI 10.1371/JOURNAL.PONE.0179488.

Gao W, Zheng X, Zhao S. 2021. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF. *Journal of Physics. Conference Series* **1848(1)**:012083 DOI 10.1088/1742-6596/1848/1/012083.

Güneş A, Tantuğ AC. 2018. Turkish named entity recognition with deep learning. In: *26th IEEE signal processing and communications applications conference (SIU)*. 1–4 DOI 10.1109/SIU.2018.8404500.

Hakala K, Pyysalo S. 2019. Biomedical named entity recognition with multilingual BERT. In: *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*. 56–61 DOI 10.18653/V1/D19-5709.

Hu Y, Nuo M, Tang C. 2019. A deep learning approach for chinese tourism field attribute extraction. In: *Proceedings—2019 15th international conference on computational intelligence and security, CIS 2019*. 108–112 DOI 10.1109/CIS.2019.00031.

Jwa M-C, Jwa J-W. 2022. Development of tourism information named entity recognition datasets for the fine-tune KoBERT-CRF model. *International Journal of Internet, Broadcasting and Communication* **14(2)**:55–62 DOI 10.7236/IJIBC.2022.14.2.55.

Kaur D, Gupta V. 2020. A survey of named entity recognition in English and other Indian languages. *International Journal of Computer Science Issues—IJCSI* **7(6)**:2010.

Labusch K, Neudecker C, Zellhöfer D. 2022. BERT for named entity recognition in contemporary and historical German. *Available at https://digital.staatsbibliothek-berlin.de/oai* (accessed on 27 March 2022).

Mansouri A, Affendey LS, Mamat A. 2008. Named Entity Recognition Approaches. *International Journal of Computer Science and Network Security* **8(2)**:339.

Rani S, Arifin DM, Huda SN, Fudholi DH. 2021. Case-based mobile tourism attractions recommender system. *IOP Conference Series: Materials Science and Engineering* **1077(1)**:012009 DOI 10.1088/1757-899x/1077/1/012009.

Saputro KE, Kusumawardani SS, Fauziati S. 2016. Development of semi-supervised named entity recognition to discover new tourism places. In: *Proceedings—2016 2nd international conference on science and technology-computer, ICST 2016*. 124–128 DOI 10.1109/ICSTC.2016.7877360.

Schofield P. 2017. Holiday destination chosen based on how 'Instagrammable' the holiday pics will be. Schofields Ltd. *Available at https://www.schofields.ltd.uk/blog/5123/two-fifths-of-millennials-choose-their-holiday- destination-based-on-how-instagrammable-the-holiday-pics-will-be/*.

Seok M, Song H-J, Park C-Y, Kim J-D, Kim Y-S. 2016. Named entity recognition using word embedding as a feature. *The International Journal of Software Engineering & Applications (IJSEA)* **10(2)**:93–104 DOI 10.14257/ijseia.2016.10.2.08.

Sobhana N, Mitra P, Ghosh SK. 2010. Conditional random field based named entity recognition in geological text. *IJCA—International Journal of Computer Applications* **1(3)**:143–147 DOI 10.5120/72-166.

Souza F, Nogueira R, Lotufo R. 2019. Portuguese named entity recognition using BERT-CRF. Accessed: May 25, 2022. ArXiv arXiv:1909.10649.

Swarbrooke J, Page SJ. 2002. *Development and management of visitor attractions*. Routledge: Butterworth-Heinemann DOI 10.1016/B978-0-7506-5169-1.50050-9.

Taher E, Hoseini SA, Shamsfard M. 2020. Beheshti-NER: persian named entity recognition using BERT. ArXiv arXiv:2003.08875.

UNWTO. 2020. UNWTO world tourism barometer and statistical annex, 2020. *UNWTO World Tourism Barometer* **18(1)**:1–48 DOI 10.18111/wtobarometereng.2020.18.1.1.

UNWTO. 2022. Tourism grows 4% in 2021 but remains far below pre-pandemic levels. *Available at https://www.unwto.org/news/tourism-grows-4-in-2021-but-remains-far-below-pre-pandemic-levels* (accessed on 28 March 2022).

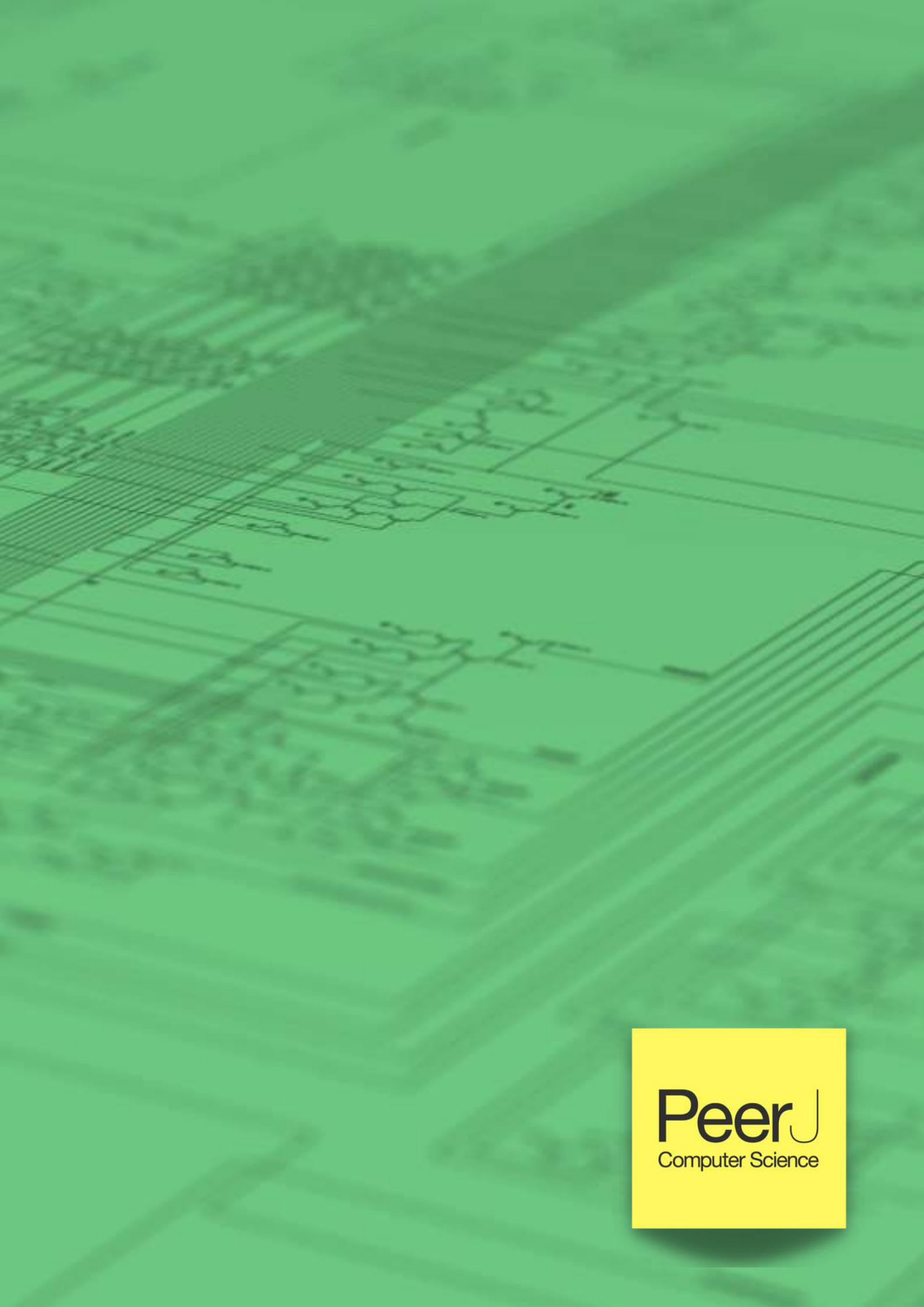
Vijayakrishna R, Sobha L. 2008. Domain focused named entity recognizer for tamil using conditional random fields. In: *Proceedings of the IJCNLP 2008 Workshop on NER for South and South East Asian Languages. January*. 59–66.

Xue L, Cao H, Ye F, Qin Y. 2019. A method of chinese tourism named entity recognition based on bblc model. In: *Proceedings—2019 IEEE SmartWorld, ubiquitous intelligence and computing, advanced and trusted computing, scalable computing and communications, Internet of People and smart city innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*. 1722–1727 DOI 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00307.

Zahra A, Hidayatullah AF, Rani S. 2022. Bidirectional long-short term memory and conditional random field for tourism named entity recognition. *IAES International Journal of Artificial Intelligence (IJ-AI)* **11(4)**:1270 DOI 10.11591/ijai.v11.i4.pp1270-1277.

Zeng D, Sun C, Lin L, Liu B. 2017. LSTM-CRF for drug-named entity recognition. *Entropy* **19(6)**:283 DOI 10.3390/e19060283.

Zhuang H, Wang F, Bo S, Huang Y. 2021. A BERT based chinese named entity recognition method on ASEAN news. *Journal of Physics. Conference Series* **1848(1)**:012101 DOI 10.1088/1742-6596/1848/1/012101.



PeerJ
Computer Science