UNIVERSIDAD INTERNACIONAL SEK

DIGITAL SCHOOL

Trabajo de fin de carrera titulado:

Revisión sistemática y análisis de metodologías qué utilizan técnicas de minería de datos y aprendizaje automático para detección del *trolling* en las redes sociales

Realizado por:

Hugo Patricio Segovia Granda

Director del proyecto:

PhD. Joe Carrión Jumbo

Como requisito para la obtención del título de

MAGÍSTER EN SISTEMAS DE INFORMACIÓN MENCIÓN EN DATA SCIENCE

Quito, octubre del 2022

DECLARACION JURAMENTADA

Yo, **HUGO PATRICIO SEGOVIA GRANDA**, con cédula de identidad **0502213952**, declaro bajo juramento que el trabajo aquí desarrollado es de mi autoría, que no ha sido previamente presentado para ningún grado a calificación profesional; y, que ha consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración, cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la UNIVERSIDAD INTERNACIONAL SEK, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normativa institucional vigente.

HUGO PATRICIO SEGOVIA GRANDA

C.C: 0502213952

DECLARATORIA

El presente trabajo de investigación titulado:

Revisión sistémica y análisis de metodologías qué utilizan técnicas de minería de datos y aprendizaje automático para detección del *trolling* en las redes sociales

Realizado por:

HUGO PATRICIO SEGOVIA GRANDA

Como requisito para la Obtención del Título de:

MÁSTER EN SISTEMAS DE INFORMACION MENCION EN DATA SCIENCE

Ha sido dirigido por el profesor

PhD. JOE CARRION JUMBO

Quien considera que constituye un trabajo original de su autor

PhD. Joe Carrión Jumbo,

DIRECTOR

PROFESORES INFORMANTES

Después de revisar el trabajo presentado, lo ha calificado como apto para su defensa oral ante el tribunal examinador.

PhD. Diego Riofrio Luzcano

PhD. Silvia Martínez Mosquera

Quito, Octubre 2022

Dedicatoria

Quiero dedicar este trabajo a mi querida esposa Lorena, a mis hijos Karolina y Pedro José, mis padres, Hugo y Amanda, a mis hermanas, Miryam, Gissela, María Augusta y Gabriela, quienes permanentemente han estado apoyándome en todo momento. Gracias a ustedes mi querida familia, de mi parte seguiré preparándome en la vida para ser un mejor ser humano y un profesional que aporte con bien a la sociedad.

Agradecimiento

Agradezco a toda mi familia por el apoyo brindado, especialmente a mi esposa y mis hijos que han sido inspiración para mi superación.

Un agradecimiento formal a la Universidad Internacional SEK y a mi tutor el PhD. Joe Carrión Jumbo, quienes han guiado cada uno de los pasos en este proceso de formación.

Debo agradecer también y de manera especial a la empresa Procesadora Nacional de Alimentos "PRONACA" quién ha apoyado mi trayectoria profesional.

Resumen

El desarrollo de las redes sociales digitales ha facilitado a las personas mejorar su relacionamiento pues han permitido compartir por estos medios historias y experiencias expresadas en frases, imágenes, videos, audios, otros, toda esta información liberada dentro de grupos virtuales; sin embargo, no todo ha sido color de rosa ya que han existido también comportamientos anómalos en estos grupos y perfiles de usuarios que incitan a la violencia, el acoso, la burla, o la desinformación; a este tipo de comportamiento se lo denomina "*trolling*".

Las formas de expresar el *trolling* pueden ser variadas y van desde una muestra de irrespeto, sarcasmo, cyberbulling o influencia negativa hacia un grupo social o una persona, pudiendo ocasionar en algunos seres humanos problemas psicológicos complejos incluso suicidios en los casos más complicados.

Para contrarrestar esta realidad muchos investigadores han desarrollado estudios que permiten identificar el trolling en las redes sociales mediante técnicas de minería de datos, software especializado o inteligencia artificial. El resultado de este trabajo permite a otros investigadores identificar técnicas, modelos, métodos y algoritmos utilizados para la detección del *trolling*.

Es este trabajo se logró determinar los algoritmos usados más frecuentemente para este fin, entre estos se nombran a Random Forest, Regression Lineal, Support Vector Machine, Naïve Bayes, Kmeans, SMO, Gradient Boost, CNN. Algo que es preciso resaltar es la complejidad que tuvieron algunos investigadores a la hora de recabar datos de fuentes para su análisis, ya que la información no es libre, sin embargo, se logró identificar que la fuente de datos más usada pasa el análisis fue proveniente de la red social Twitter.

También se logró concluir que existe un avance significativo en la identificación del *trolling* mediante técnicas de minería de datos y aprendizaje automático, no obstante se observa que ningún modelo entrega un alto porcentaje de certeza en la clasificación del contenido de tipo *trolling* y para la detección de perfiles de tipo *troll*, quizá la poca la poca información que entregan las empresas que son dueñas de los servicios de redes sociales sea la causa, así como las diferentes expresiones semánticas que tiene el lenguaje natural en diferentes culturas y países.

Palabras clave: aprendizaje de máquina, algoritmos, trolling, ciberacoso, redes sociales.

Abstract

The development of digital social networks has made it easier for people to improve their relationships as they have allowed them to share stories and experiences which are expressed in phrases, images, videos, audios, etc. All this information has been shared within virtual groups. However, not everything is all rosy, there are also abnormal behaviors in these groups and user profiles that incite violence, harassment, mockery or misinformation. This type of conduct is called "*trolling*".

Trolling can be expressed in many ways, such as, disrespect, sarcasm, cyberbullying or showing a negative influence on a group of people or an individual. This could cause psychological problems, which could even lead to suicide.

To counter this reality, many researchers have developed studies that allow trolling to be identified through data mining techniques in social media, specialized software or artificial intelligence. The result of this work allows other researchers to identify techniques, models, methods and algorithms used for trolling detection.

In this research, it was possible to determine the most frequently used algorithms for this purpose, among these are Random Forest, Linear Regression, Support Vector Machine, Naïve Bayes, Kmeans, SMO, Gradient Boost and CNN. Something that must be highlighted is the complexity that some researchers had when collecting data from sources for analysis, since the information is not free, however, it was possible to identify that the most used data source for the analysis was extracted from Twitter.

It was also possible to conclude that there is a significant progress in the identification of trolling through data mining and machine learning techniques, however it is observed that no model provides a high percentage of certainty in the classification of trolling-type content and for the detection of trolling and troll-type profiles. It could be the limited information provided by the companies that own the social network services., as well as the different semantic expressions that natural language has in different cultures and countries.

Keywords: machine learning, algorithms, trolling, cyberbulling, social networks.

Tabla de Contenidos

CAPÍTULO 1 12
INTRODUCCIÓN 12
1.1 Planteamiento del problema 12
1.2 Justificación 13
1.3 Objetivos 15
1.3.1 Objetivo General 15
1.3.2 Objetivos Específicos 15
1.4 Marco teórico 15
1.4.1 La minería de datos 17
1.4.2 Aprendizaje supervisado 17
1.4.3 Aprendizaje no supervisado 18
1.4.4 Aprendizaje automático 18
1.4.6 Deep learning 19
1.4.7 Las redes sociales digitales 19
1.4.8 Los *Trolls* y el *trolling* 20
CAPÍTULO 2 22
METODO 22
2.1 Planificación 24
2.2 Búsqueda e identificación de fuentes de información 26
2.3 Relevamiento y conocimiento de la información 27
2.4 Análisis de la información 31
2.5 Hallazgos 33
CAPÍTULO 3 33
RESULTADOS 34
3.1 Análisis de los modelos utilizados para la detección del *trolling* 34
3.2 Aplicaciones desarrolladas para detectar el *trolling* 78
3.3 Análisis del *trolling* mediante el reconocimiento de imágenes 87
3.4 Detección del *trolling* mediante tensores 88
3.5 Resumen de los estudios y la efectividad de los modelos y algoritmos usados 89
3.6 Resumen de las aplicaciones desarrolladas 101
CAPÍTULO 4 102

CONCLUSIONES Y TRABAJOS FUTUROS 103
BIBLIOGRAFÍA 105
ANEXOS 111

Lista de Figuras

Figura 1 23
Figura 2 42
Figura 3 42
Figura 4 44
Figura 5 45
Figura 6 46
Figura 7 49
Figura 8 50
Figura 9 51
Figura 10 51
Figura 11 54
Figura 12 56
Figura 13 57
Figura 14 61
Figura 15 63
Figura 16 64
Figura 17 65
Figura 18 69
Figura 19 70
Figura 20 71
Figura 21 71
Figura 22 72
Figura 23 73
Figura 24 75
Figura 25 80
Figura 26 82
Figura 27 83
Figura 28 86
Figura 29 86
Figura 30 88

Lista de Tablas

Tabla 1 25
Tabla 2 25
Tabla 3 25
Tabla 4 26
Tabla 5 27
Tabla 6 28
Tabla 7 29
Tabla 8 30
Tabla 9 30
Tabla 10 31
Tabla 11 31
Tabla 12 32
Tabla 13 35
Tabla 14 40
Tabla 15 41
Tabla 16 43
Tabla 17 46
Tabla 18 48
Tabla 19 52
Tabla 20 53
Tabla 21 54
Tabla 22 55
Tabla 23 58
Tabla 24 62
Tabla 25 65
Tabla 26 66
Tabla 27 67
Tabla 28 67
Tabla 29 69
Tabla 30 74
Tabla 31 76
Tabla 32 78
Tabla 33 80
Tabla 34 81
Tabla 35 84
Tabla 36 87
Tabla 37 89
Tabla 38 96
Tabla 42 101
Tabla 40 111

CAPÍTULO 1

INTRODUCCIÓN

El realizar un análisis sistemático de los estudios identificados en el mundo sobre alguna temática en especial, abre las posibilidades de mayor conocimiento hacia los investigadores, adicionalmente que ayudará a la no duplicación de esfuerzos realizados (Tebes et al., 2019).

Este capítulo detalla la situación actual entendida como problemática, que en su generalidad se interpreta como no disponer de información consolida y analizada para la identificación del trolling mediante técnicas de minería de datos.

Todo lo que se detalla a continuación se sustenta en conceptos relacionados a la mimería de datos y al aprendizaje automático, también parte de una base metodológica.

1.1 Planteamiento del problema

La información que se genera en las redes sociales ha crecido significativamente, y se prevé que en los siguientes años esta tendencia se desarrolle aún más. En relación con este crecimiento el mundo ha experimentado el nacimiento de plataformas y servicios tecnológicos que los seres humanos utilizan para relacionarse, esto ha dado origen a las redes sociales.

A lo largo de este estudio se presentan varias tendencias de comportamientos humanos que son ventilados en las redes sociales no solamente para compartir experiencias, estilos de vida o algún tipo de criterio sino también para causar caos, burla o desacreditación. El identificar la forma de contrarrestar este tipo de comportamientos es un reto muy importante que algunos investigadores se han propuesto, por consiguiente, el conocer las técnicas y resultados de estudios que se han propuesto como objetivo el usar la minería de datos y el aprendizaje automático para reconocer y mitigar estas acciones denominadas *trolling* o ciberacoso será un objetivo importante.

Entonces, se plantea una interrogante ¿cómo determinar el número de trabajos que se han desarrollado en el mundo y a lo largo del tiempo relacionados a la identificación del *trolling* mediante técnicas de minería de datos, inteligencia artificial y aprendizaje automático?, estos cuestionamientos son necesarios para marcar un estado del arte actual sobre las técnicas, modelos y algoritmos utilizados, las fuentes de datos y sus resultados (Tebes et al., 2019).

Lo que se conoce y se mide se puede mejorar, por tanto, el no disponer de una revisión sistemática que visualice e identifique aquellas técnicas utilizadas y sus resultados probablemente haga que muchos trabajos relacionados al tema abordado en este estudio sigan repitiendo los mismos análisis con los mismos métodos y las mismas fuentes de datos, y quizá los mismos resultados, lo que significa que los avances no son medidos o por menos identificados y el resultado sigue siendo el mismo o con pocas variaciones, por tal motivo la revisión sistemática es una metodología que es utilizada para obtener evidencia de artículos científicos que son almacenados el repositorios digitales (Tebes et al., 2019).

1.2 Justificación

Según el portal Statista, 2022 especializado en estadísticas para datos de mercado, investigaciones de mercado y estudios de mercado, indica entre uno de sus estudios que la penetración de Internet en varios continentes desde el 2009 hasta el 2019 ha variado significativamente con un crecimiento importante llegando hacer en algunos casos hasta de un 80% mayor, esto significa que muchas personas en el mundo tienen acceso a los servicios de Internet, cada vez más se puede acceder a información de texto, audio, gráfica, y videos disponibles en el medio. El aprovechamiento de este recurso de comunicación ha dado lugar al crecimiento de páginas de servicios web que permiten la interacción de personas, grupos, y comunidades, como foros, sitios de noticias, redes sociales, entre algunos.

Se toma como base las redes sociales, estos son medios que permiten en su funcionalidad la interacción con otras personas o grupos de personas mediate la publicación de textos, fotografías, o videos. existen dos tipos de redes sociales “publicas” y “privadas”; este estudio se centra en el análisis al contenido de mensajes en sitios web y redes sociales públicas, a continuación, se nombran algunas de las más importantes, como lo son Facebook, Instagram, Twitter, LinkedIn WhatsApp, Snapchat, Google+, Pinterest. Para hacer uso de estas redes de servicios es necesario ser miembros y para el inicio de la membresía se realiza el ingreso de nuestros datos personales en una forma electrónica, una vez aceptada la membresía se habilita inmediatamente el ser parte de ellas, y así consumir y publicar cualquier tipo de información (Ceballos Saavedra, 2014).

Las discusiones en redes sociales son variadas, el enfoque de aprobación o disgusto de un tema en particular en un grupo de usuarios va desde un simple like o pulgar levantado sinónimo de

aprobación hasta una conversación controversial entre los usuarios, con quienes se puede intercambiar diferentes puntos de vista sobre alguna temática expuesta en este medio, se ha identificado que algunos comentarios generados en estos medios podían estar fuera del contexto moral, político o que inciten algún tipo de agresión o acoso, que en muchas ocasiones ha desnudado los valores morales de las personas (Basalingappa et al., 2019).

Una de las preguntas que podría rondar la mente de los usuarios es ¿Cómo saber si la persona con quién estamos interactuando en una de estas redes sociales es quien dice que es? Además, surgen otras preguntas como ¿Podrían existir perfiles de usuarios falsos en las redes sociales? A lo largo de este trabajo se explican los hallazgos que sustentan esta repuesta por lo que una persona que ingresa a una red social podría falsificar fácilmente una identidad, sobre la falsificación se comenta que existen en Internet varios servicios tecnológicos que usan inteligencia artificial para generar automáticamente una fotografía de una persona que no existe en el mundo (Kashish Hora, n.d.).

Entonces, si se considera que existe la posibilidad de falsificar un perfil en una red social, nos preguntamos nuevamente ¿existe la posibilidad de crear perfiles de personas en redes sociales de forma automática y que estos interactúen con otros usuarios reales de forma programada?, es decir sin la intervención humana. Hofmann et al., (2020) comentan que existe la posibilidad de disponer de automatismos que creen de forma consecutiva perfiles de redes sociales con nombres y fotografías de personas ficticias y los cuales luego sean utilizados para diferentes acciones de comunicación.

En este sentido, ¿se asume la posibilidad de que existan perfiles falsos creados?; Bishop (2013) expone la posibilidad de tener un ejército de personas virtuales atrás de una red social que interactúa con perfiles de personas reales, pero ¿cuál es el propósito de este comportamiento? ¿para qué necesito un perfil falso?, los perfiles falsos por lo general están siendo utilizados sesgar comportamientos de grupos humanos con desinformación, incidir en cambios de tendencias, causar caos mediante la intimidación, el insulto o simplemente la contradicción.

Para comprender toda esta realidad presentada se realiza un análisis sistemático a varios estudios que tratan de identificar los perfiles denominados acosadores o *Trolls.* Algunos cuestionamientos surgen como por ejemplo ¿qué tecnología usan? o ¿cuál es el mecanismo de identificación?, en el desarrollo de este trabajo se abordarán las técnicas y conceptos utilizados.

El determinar una fecha de inicial para recopilar y revisar la información sobre los estudios realizados relacionados al tema es indispensable, por lo tanto, se establece que la fecha de inicio de la búsqueda es el año 2001 marcando así un rango de aproximadamente 20 años hasta la fecha actual en que se comienza el análisis sistemático, octubre del año 2021; estas fechas son establecidas en base al análisis realizado por Ceballos Saavedra (2014) quien comenta la aparición de los sitios web en los que se promociona las redes de círculos de amigos desde el año 2002 por lo tanto este inicio nos entrega una guía del tiempo de búsqueda de los estudios relacionados.

1.3 Objetivos

1.3.1 Objetivo General

Realizar una revisión sistemática de las metodologías de minería de datos y aprendizaje automático utilizadas en la detección del *trolling* desde el inicio del año 2001 hasta el mes de octubre del año 2021, mediante la aplicación de métodos y técnicas de búsqueda, organización y síntesis de los documentos publicados en las fuentes de consulta reconocidas en el mundo.

1.3.2 Objetivos Específicos

- Identificar la existencia de revisiones sistemáticas relacionadas al tema propuesto, mediante la búsqueda y análisis de los estudios digitales localizados en bases de datos documentales.
- Realizar la clasificación de los estudios existentes en el mundo relacionados al tema propuesto, a través de la observación a los trabajos recopilados obteniendo así una visión general de sus aplicaciones.
- Determinar los modelos analíticos y tecnológicos utilizados en los estudios realizados, utilizando una matriz que permita la recopilación, clasificación y visualización de forma ágil.
- Evidenciar las tendencias de uso de los algoritmos utilizados y sus resultados, a través de la identificación de los modelos aplicados en los estudios.
- Determinar cuáles fueron las técnicas y fuentes de datos mayormente utilizadas en el desarrollo de los estudios analizados.

1.4 Marco teórico

El sustento metodológico está presente en las investigaciones científicas pues no solo aporta un marco de referencia sobre la temática a analizar, sino que también determina una línea base de conocimientos que se utilizarán a lo largo del desarrollo del trabajo de investigación. En esta

sección se sustentan los conocimientos base que se utilizan en el análisis sistemático de los estudios recabados.

La ciencia de datos y la minería de datos son terminologías que se podría decir que no son relativamente nuevas, estas utilizan varias técnicas de análisis de datos y formas de extraer la información así como análisis predictivos identificando patrones de datos que pueden ser usados (Orallo et al., 2004). El conocimiento estadístico – tecnológico apoya al desarrollo de la disciplina, aporta una comprensión profunda de los algoritmos matemáticos que son usados para propósitos probabilísticos y para la gestión de las bases de datos usadas para el almacenamiento, los datos utilizados para los análisis son un insumo importante sobre los cuales se usan las técnicas de minería de datos y aprendizaje automático (Orallo et al., 2004).

Las organizaciones han brindado mayor importancia en la gestión de los datos que se generan, ya que en muchos casos los datos son considerados como parte del producto o servicio siendo estos partes de un activo organizacional, sin ellos es imposible desarrollar un producto o servicio. Los datos se vuelven una memoria organizacional, con ellos se puede explicar el pasado, entender el presente y predecir el futuro (Orallo et al., 2004).

Para empresas como Facebook (Mark Zuckerberg, 2004), Instagram (Kevin Systrom, Mike Krieger (Burbn, 2010), y Twitter (Jack Dorsey, Noah Glass, Biz Stone, 2006), el generar datos y guardarlos para que posteriormente sean analizados mediante técnicas de analítica, esto ha impulsado su desarrollo significativamente gracias a que pueden predecir comportamientos no solo unipersonales sino también colectivos incrementando así el conocimiento de la información sobre los grupos humanos.

Entonces, se considera que para analizar datos y poder pronosticar alguna tendencia o comportamiento de los mismos es necesario tener conocimientos específicos en estadística y tecnología de la información para poder así aplicar modelos predictivos, de aprendizaje automático, o de aprendizaje profundo, términos que también se abordarán en esta revisión sistemática (Orallo et al., 2004).

1.4.1 La minería de datos

La minería de datos se sustenta en producir nuevo conocimiento que se pueda utilizar mediante la aplicación de un modelo basado en datos que han sido recopilados para este fin, el modelo describe lo patrones relacionados a los datos que posteriormente son usados para realizar pronósticos o predicciones que ayudan a comprender el posible escenario de comportamiento futuro (Orallo et al., 2004). Los modelos pueden ser descriptivos y predictivos; en los descriptivos se identifican patrones que pueden explicar y resumir los conjuntos de datos; mientras que los predictivos pueden estimar valores futuros a partir de variables de interés (Orallo et al., 2004).

Ya en el campo de la mimería predictiva se tiene modelos como los de clasificación o los de regresión los cuales usan diferentes algoritmos para el procesamiento de los datos (Orallo et al., 2004).

No solamente el conocer cómo funcionan los algoritmos es suficiente para desarrollar un modelo es importante también conocer cómo se preparan los datos desde la selección de fuentes hasta su resultado. Las características o fases utilizadas en la minería de datos las describen Orallo et al. (2004), quienes sintetizan en las siguientes acciones.

1. Integración y recopilación.
2. Selección limpieza y transformación.
3. Minería de datos.
4. Evaluación e interpretación.
5. Difusión y uso.

Cada una de las etapas una de las etapas juega un papel importante en el proceso de minería de datos para la obtención final de conocimiento.

1.4.2 Aprendizaje supervisado

El aprendizaje supervisado comienza desde la selección del conjunto de datos y una visión de cómo se clasificarán estos. El objetivo es encontrar el patrón que se pude aplicar en este proceso analítico, en un ejemplo práctico podría ser el crear modelo de clasificación que distinga de millones de fotografías solo aquellas que sean solamente de animales, basándose en imágenes ya descritas (Moreno et al., 2004).

El aprendizaje supervisado se clasifica en modelos de regresión y modelos de clasificación. Los modelos de regresión se basan en predecir valores continuos a partir de la información histórica,

por ejemplo, las predicciones meteorológicas. Mientras que en los de clasificación el objetivo es agrupar los datos de acuerdo con los datos históricos previamente etiquetados, por ejemplo, detecciones de fraude de identidad (Orallo et al., 2004).

1.4.3 Aprendizaje no supervisado

El aprendizaje no supervisado es utilizado cuando el problema ha requerido de una gran cantidad de datos que no disponen de etiquetas, por ejemplo en el mundo de las redes sociales existen servicios como Facebook, Instagram, Twitter, que disponen de grandes cantidades de datos sin etiquetas, comprender esta gran cantidad es datos requiere la intervención de algoritmos que clasifican los datos basados en patrones; entonces el aprendizaje no supervisado dirige procesos autónomos que analizan datos sin intervención humana, por ejemplo para identificar patrones de perfiles los cuales estén interactuando en las redes insertando datos no autorizados.

Otro ejemplo de aprendizaje no supervisado es el que se utiliza para detectar virus en los sistemas informáticos, cada vez más estos sistemas clasifican la información que ingresa a una red informática, bloqueando o alertando a sus usuarios de anomalías similar comportamiento tienen los datos *outliers* que se separan significativamente de los datos de tendencia central (Orallo et al., 2004).

1.4.4 Aprendizaje automático

El aprendizaje automático llamado en el argot de la analítica de datos como *Machine Learning* (ML) es capaz de ejecutar algoritmos y programas que son capaces de aprender y que en conjunto con los análisis estadísticos constituyen el corazón del análisis de datos inteligente, el dotar de capacidad o inteligencia de **razonamiento a una máquina** ha sido el reto de muchas empresas, un ejemplo claro de esto es Netflix que a enero del 2021 ha superado los 20 millones de usuarios (Abigail Orús, 2022). Hoy en día el motor de búsqueda de sus películas o series se basa en aprendizaje automático, es decir comprende nuestros comportamientos y nos sugiere películas que pueden interesar a sus suscriptores, ¿Qué hay detrás de esto?, la respuesta es algoritmos que están ejecutándose, recogiendo información, aprendiendo los intereses de las personas y entregando sus análisis (Orallo et al., 2004).

El *Machine Learning* es un concepto que permite a los sistemas aprender a partir del análisis y el procesamiento de datos, en lugar de realizar tareas o funciones de programación convencional, el ML logra tener resultados mediante el uso de algoritmos.

1.4.6 Deep learning

El *Deep Learning* es una técnica de aprendizaje automático (ML) basado en modelos de redes neuronales que analizan los datos en forma de capas sucesivas para aprender de los datos de forma repetitiva, es muy útil esta técnica para procesar datos sin una estructura definida o datos no estructurados; estas redes neuronales emulan el cerebro humano para identificar patrones de imágenes que ayuden a asociar respuestas, por ejemplo para un infante de 5 años es fácil distinguir la cara de su madre o de su maestro pero para los sistemas computacionales es todo un reto debido a esto el aporte de las redes neuronales y *Deep Learning* ha sido valioso para el reconocimiento de imágenes (Orallo et al., 2004).

1.4.7 Las redes sociales digitales

Las redes sociales digitales (RSD) son servicios computacionales basados principalmente en plataformas Web que permiten a las personas o empresas construir un perfil digital dentro del sistema, estos perfiles y su contenido a su vez sean compartidos con otros usuarios de la red. El nacimiento de las RSD se remonta a los años 1995 cuando Randy Conrads crea el sitio web classmates.com, su objetivo fue que muchas personas mantengan el contacto con sus compañeros del colegio o de la universidad por medio de su sistema (Uribe-saavedra, 2011).

Dasilva et al. (2013) comentan en su análisis que las redes sociales tienen una influencia importante en las estrategias de comunicación, es por eso por lo que muchas empresas se han visto obligadas a integrarlas en sus ecosistemas.

El éxito de las redes sociales ha sido tan amplio que ha llevado a ser uno de los principales medios de consumo de contenido. Se ha identificado que los internautas no consumen información únicamente de los buscadores, sino que también usan las recomendaciones y *links.* En septiembre el año 2010 Facebook (Mark Zuckerberg, 2004) era considerada como la segunda fuente de tráfico para gran parte de los sitios de contenido en España; detrás de Google (Dasilva et al., 2013).

Dasilva et al. (2013) exponen en su trabajo dos de las principales marcas comerciales de redes como Facebook (Mark Zuckerberg, 2004) y Twitter (Jack Dorsey, Noah Glass, Biz Stone, 2006), estas están presentes en la vida de muchas personas alrededor del mundo las cuales albergan contenido de, usuarios, productos, noticias, tendencias, política, lo que los hace atractivo para las empresas que quieren promocionar sus productos o servicios (Dasilva et al., 2013).

Por lo tanto, las redes sociales se consideran un medio para difundir información de manera directa y eficaz, pero en muchos casos sin control, lo que significa que cualquier perfil puede publicar un comentario u opinión

1.4.8 Los *Trolls* y el *trolling*

El Internet se ha convertido un medio de comunicación globalizado, mediante el cual se puede acceder a información de caracteristicas profesionales, negocios, ocio. Algunas estadísticas indican que a finales del año 2019 el 53.6% de la población mundial accedió diariamente a Internet e hizo uso principalmente de las redes sociales como Facebook, Twitter, Instagram entre otras (Lupano Perugini et al., 2021).

Se pudo determinar que en un estudio realizado en Argentina mediante una técnica denominada análisis factorial se aislaron los tres motivos por los cuales las personas ingresan a una red social, esto son: contactos con amigos y familiares, ocio (pasatiempo y exhibicionismo) y el último para conocer nuevas personas y relacionarse con ellas (Lupano Perugini et al., 2021).

Este uso de las redes sociales ha traído beneficios comunicacionales importantes pues debido a esto se ha incrementado las posibilidades de relacionamiento humano, sin embargo, no todo es bueno ya que se ha desarrollado también comportamientos antisociales denominados *trolling.* Se ha visto que estos comportamientos pueden causar en las víctimas ansiedad, depresión o bajos niveles de autoestima (Lupano Perugini et al., 2021).

Los comentarios acosadores, causantes de caos son perpetrados por otros usuarios a los que se los denomina *trolls,* estos pueden actuar de forma unipersonal o en asociaciones y pueden atacar a individuos selectivos, grupos sociales, agrupaciones políticas, empresas, entidades, corporaciones entre otros.

Herrera (2021) expone su trabajo y manifiesta que en Chile muchas figuras influyentes han acusado a las redes sociales por constituir un peligro para la democracia y caracterizan a los usuarios como intolerantes y destructivos.

Existen algunas definiciones para los *trols* por ejemplo se los define como acosadores, abusadores términos equiparados a la acción del *troleo* como ciberacoso. Los objetivos de los *trols* son variados como callar al usuario oponente y desvirtuar su contenido evitando que siga publicando comentarios (Herrera, 2021).

CAPÍTULO 2
METODO

Para la realización de este trabajo se toma como referencia el estudio de Tebes et al. (2019) quienes exponen en su trabajo que en el año 2004 la investigadora Barbara Kitchenham publica un informe técnico sobre las revisiones sistemáticas de literatura (RSL) basadas en investigaciones realizadas al dominio de la medicina, considerando a este autor como el precursor de esta metodología de investigación. En los siguientes años las RSL fueron adoptadas también por las ciencias de ingeniería y software para el análisis de estudios primarios y secundarios.

La metodología plantea tres fases definidas de abordaje para realizar el análisis sistemático de literatura.

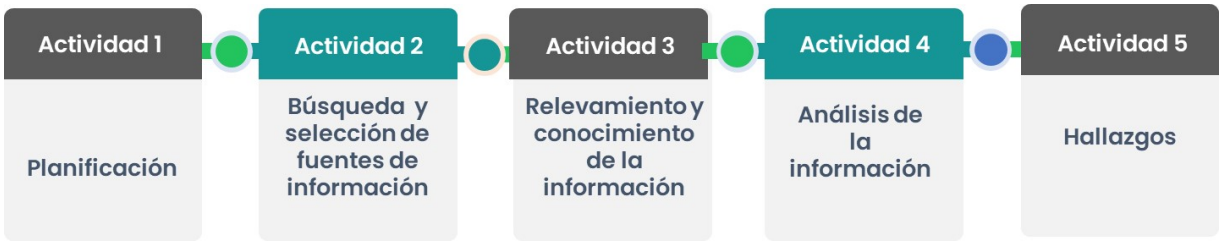
1. ***Fase 1: Planificar la revisión***
    a. Especificar preguntas para la investigación.
    b. Desarrollar un protocolo de revisión.
    c. Validar el protocolo de revisión.
2. ***Fase 2: Conducir la revisión***
    a. Identificar las fuentes de información.
    b. Seleccionar los estudios primarios.
    c. Analizar y avaluar la calidad de los estudios.
    d. Extraer los datos requeridos.
    e. Sintetizar los datos.
3. Fase 3: ***Documentar la revisión***
    a. Escribir el informe de revisión.
    b. Validar el informe.

Algo relevante y que persigue este modelo es el facilitar el entendimiento y la comunicación para la comunidad interesada.

En este sentido y para lograr el objetivo del análisis sistemático en este trabajo se propone una metodología nueva que consta de 5 actividades, para así conseguir un abordaje eficiente sobre un proyecto de análisis sistemático de la literatura científica; para ello se ha toma como base la metodología de Kitchenham (Tebes et al., 2019). Cada una de las actividades desarrolladas entrega información a la siguiente para tener una secuencia y un rápido resultado. Las actividades son pasos secuenciales que se deben realizar, similares a las actividades un proceso como se visualiza en la Figura 1.

Figura 1

Proceso para el análisis sistemático de revisiones de literatura.



Nota. Diagrama del proceso propuesto para la realización del análisis sistemático de literatura científica. Fuente Tebes et al. (2019).

1. ***Actividad 1: Planificación***
    a. Visión, objetivos y metas.
    b. Obtención de recursos.
2. ***Actividad 2: Búsqueda y selección de fuentes de información***
    a. Identificación, análisis y selección de fuentes.
3. ***Actividad 3: Relevamiento y conocimiento de la información***
    a. Validación de la información.
4. ***Actividad 4: Análisis de la Información***
    a. Analizar los estudios.
    b. Extraer la información relevante.

c. Sintetizar información.

5. ***Actividad 5: Hallazgos***

a. Desarrollar el informe.

b. Validar el informe.

El desarrollo de este estudio inicia con el planteamiento de un objetivo alcanzable, realizable y viable más una planificación de varias actividades que componen el proceso antes mencionado, estas guiarán el desarrollo del documento.

Como se ha mencionado, el proceso planteado tiene 5 actividades, cada una con un nivel de desarrollo y profundidad importante. En cada actividad se realizan acciones únicas y con un tiempo establecido para llegar a cumplir los objetivos, a continuación, se describirán cada una de ellas.

2.1 Planificación

El análisis sistemático inicia desde la planificación, en esta etapa se realiza un ejercicio de visión del tema propuesto a analizar, así como el planteamiento de los objetivos y metas, se plantean los recursos necesarios.

En resumen, la planificación proyecta las expectativas del análisis futuro donde se establecen objetivos claros y alcanzables, adicionalmente se busca identificar todos los recursos necesarios para desarrollar el análisis sistemático de literatura. Es importante que al plantearse el objetivo se contesten algunas preguntas que ayudarán a comprender la razón, la forma de abordaje y los recursos necesarios, estas preguntas son ¿Por qué?, ¿Cuáles? ¿Cómo? y ¿Con qué?

La primera actividad es el planteamiento de los objetivos, para conseguir este resultado se busca inicialmente imaginar el **por qué** se está realizando este estudio desde la perspectiva de agregación de valor, cuáles son los motivadores que promueven este análisis e inclusive cuál es el universo de personas que pueden consumir esta información.
El análisis sistemático de los estudios enfocados a la identificación del *trolling* es el objetivo planteado, para comprender un poco más el desarrollo de esta etapa se detalla en la Tabla 1 los motivadores planteados.

Tabla 1

Definir el objeto de la investigación

Motivadores para establecer el objetivo: Por qué
1. ¿Por qué es importante diagnosticar el *trolling*?
2. ¿Cuáles son las consecuencias de esta realidad?
3. ¿Qué pasa si no se aborda esta investigación?
4. ¿Cuál es el universo de usuarios que son afectados por esta realidad?
5. Qué obtendríamos del resultado de este trabajo

Nota. Motivadores para determinar el objeto de la investigación

Ahora, el planteamiento es la necesidad de conocer cómo lo vamos a realizar el estudio, es decir las acciones que determinan la operación, aquí miramos si el objetivo es alcanzable; esto se detalla en la Tabla 2.

Tabla 2

Definir el alcance de la investigación.

Motivadores para establecer el objetivo: Cómo
1. ¿Cómo se realizará un análisis profundo a las investigaciones?
2. ¿Cómo analizar y conocer los algoritmos utilizados?

Nota. Motivadores que identifican el cómo se realizará el estudio.

Definidos y desarrollados los motivadores se determinan los recursos necesarios que se utilizarán en el desarrollo, como por ejemplo el conocimiento, el tiempo y las herramientas que se disponen para este proyecto, entre algunos. Para logar concluir esta actividad se establece otro conjunto de preguntas que deben ser contestadas, las respuestas a estas preguntas identifican restricciones o riesgos que deben ser mitigados como se expone en la Tabla 3.

Tabla 3

Identificación de recursos.

Motivadores para establecer capacidades y recursos: Con qué
3.1 ¿Con qué capacidades cuento para analizar la información?
3.2 ¿Qué equipos son necesarios para este análisis y con qué dispongo?
3.3 ¿Cuál es el tiempo determinado para completar el estado del arte?

Nota. Motivadores para identificar qué recursos y capacidades son necesarias

Al concluir el desarrollo de las preguntas se obtienen más datos para comprender el alcance real del estudio planteado, estos pasos se replican también para determinar los objetivos específicos del proyecto de investigación.

2.2 Búsqueda e identificación de fuentes de información

Una vez concluida la planificación inicial, comienza la etapa a la identificación de las fuentes de información, en esta fase se buscan los repositorios y que contienen los documentos de estudios relacionados a la identificación del *trolling.*

Una lluvia de ideas marca el inicio de la actividad, en la que se listan las posibles fuentes y sitios web de información que pueden contener estudios relacionados al tema, el siguiente paso es organizar la información, para ello se detalla cada una de las fuentes y se analiza su contenido. En el sitio web de Paperpil, (2021) se publican datos sobre las bases de información académica más importantes según su área de conocimiento, estas fuentes son tomadas en cuenta para el análisis y el resultado de la selección es plasmado en la una matriz mostrada en la Tabla 4.

Tabla 4

Fuentes de información identificadas.

Repositorio	Descripción
Scopus	Base de datos de referencias bibliográficas y citas de estudios.
Latindex	Sistema de información de para revistas científicas en américa latina.
Google Schoolar	Motor de búsqueda de contenido y bibliografía científica y académica.
Redalyc	Sistema de información de revistas académicas para Latinoamérica, el caribe, España y Portugal.
Scielo	Modelo para la publicación de revistas.
IEEE Xplore	Base de datos de investigación para descubrir artículos de revistas.
ACM	Base de datos digital de información científica.
Science Direct	Bases de datos de información científica y médica.
Springer	Editorial de publicación de libros electrónicos.

Nota. Listado de fuentes de información identificadas para la realización del estudio

2.3 Relevamiento y conocimiento de la información

Establecidas las fuentes de las cuales se obtendrán los estudios, se realiza las búsquedas en cada una de ellas, usando filtros que permiten la identificación rápida de estos documentos de estudios, y para lograr una mayor efectividad se definen palabras clave que serán utilizadas en los filtros como se muestra en la Tabla 5.

Tabla 5

Listado de palabras

Palabras clave
Social network
Social networks
Facebook
Twitter
Instagram
Internet
Online
Detect
Detection
Troll
Trolls
Trolling
Model
Identification
Recognition
Catch
Perceive
Analytics
Data
Machine learning
Data mining
Survey

Nota. Listado de palabras utilizadas en los filtros de las fuentes de datos para obtener información de los estudios

Las palabras son usadas en cadenas de búsqueda en las páginas web de las fuentes de datos.

Es importante remarcar que los criterios de búsqueda para encontrar la información deben ser precisos y con un enfoque que permita encontrar la información requerida, por tanto, se expone en la Tabla 6 los ***scripts*** definidos para realizar las búsquedas en la fuente Scopus.

Tabla 6

Cadenas de búsqueda.

Repositorio	**Filtro de búsqueda**
Scopus	TITLE-ABS-KEY ((*trolls* OR *troll*) AND ("social networks" OR "social network" OR Facebook OR Twitter OR Instagram) AND (detect OR detection OR model) AND (analytics OR data OR "Machine Learning" OR "data mining") AND (“state of art" OR "systematic review" OR "state of the art”)) Script de búsqueda filtro sin estado del arte TITLE-ABS-KEY ((*trolls* OR *troll* OR *trolling*) AND ("social networks" OR "social network" OR Facebook OR Twitter OR Instagram) AND (detect OR detection OR model OR identification) AND (analytics OR data OR "Machine Learning" OR "data mining"))

Nota. Scripts utilizados en las búsquedas.

La búsqueda de estudios relacionados al tema en las diferentes fuentes arroja en algunos casos resultados similares, es decir que muchos documentos se repetían en algunas bases de datos de información científica que se analizó. Por otro lado, la fuente Google Schoolar arroja resultados que no solamente se refieren a estudios o artículos sino también a cualquier fuente independiente como una página web, un blog, una noticia, un foro; se encontraron 1950 resultados de búsqueda referentes al tema, pero con pocas posibilidades de clasificación o segmentación.

Un análisis realizado por FECYT (2017) indica los beneficios de usar Scopus, quienes comentan que esta fuente es la mayor base de datos de resúmenes y de literatura la cual dispone de varias herramientas inteligentes para analizar, controlar y visualizar artículos de investigación académica; por consiguiente luego de una revisión de los artículos encontrados en las fuentes se elige a Scopus como la fuente principal de estudios válidos que se utilizarán en el desarrollo de este proyecto como se expone en la Tabla 7.

Tabla 7

Resultados de las búsquedas

Repositorio	Tipo	Número
Scopus	Artículos	73
Latindex	Revistas	0
Google Schoolar	Buscador	1950
Redalyc	Revistas	0
Scielo	Artículos	0
IEEExplorer	Artículos	24
ACM	Artículos	27
Science Direct	Artículos	21
Springer	Artículos	2

Nota. Resultados de los artículos encontrados en las búsquedas en cada fuente de información

En la publicación de Paperpile (2021), se exponen las principales bases de datos de investigaciones académicas, clasificándolas en un ranking de 1 a 8, se menciona a Scopus como la número uno en su clasificación. Guidelines (2007) también comentan reseñas que avalan a Scopus como la fuente de información con mejor versatilidad, flexibilidad al buscar contenido, y un mayor número de estudios con validación cruzada disponibles en esta fuente, por lo tanto, se confirma nuevamente a Scopus como la base que se utilizará para obtener los estudios científicos que son analizados.

El resultado final de la búsqueda entrega **73 estudios** encontrados en la fuente de Scopus.

El entregable de esta actividad es la matriz que contiene las fuentes de datos y el número de los estudios encontrados, que se utilizan para el análisis sistemático de literatura.

Para que posteriormente se pueda analizar el conjunto de artículos encontrados con mayor precisión se realiza una clasificación previa que asegure que su contenido guarde concordancia al tema propuesto, el criterio para realizar esta clasificación se centra en una lectura rápida de cada uno de estudios encontrados pues solo ahí se podrá identificar si guardan similitud y cumplen con las características de detección del *trolling* usando técnicas de minería de datos y aprendizaje automático; este análisis se resume en la siguiente matriz de actividades realizadas, Tabla 8.

Tabla 8

Clasificación de documentos.

	Actividades
1	Ordenar los documentos por fecha.
2	Leer y analizar el resumen de cada documento.
3	Seleccionar los documentos que guardan relación.
4	Analizar el cuerpo del documento y verificar si guarda relación.
5	Seleccionar y ordenar los documentos finales que guardan relación.
6	Utilizar estos documentos para las lecturas de análisis.

Nota. Actividades para realizar la clasificación de los documentos encontrados en la fuente de información

El resultado de las preguntas anteriores debe responder claramente si el documento es o no elegible para poder usarlo.

La lectura del punto número dos nombrado en la tabla anterior arroja información adicional, para ello se usó un set de preguntas complementarias como se visualiza en la Tabla 9.

Tabla 9

Preguntas de investigación complementarias

	Preguntas
1	¿Cuántos estudios finales se encontraron?
2	¿Cuál es la clasificación de los estudios en el *ranking* de revistas Q1-Q4?

Nota. Preguntas de investigación que ayudan a determinar el número de estudios encontrados y el *ranking* en revistas

El resultado de las preguntas realizadas determina que solamente se usan 43 de los 73 estudios encontrados ya que la diferencia o sea los 30 estudios no guardan relación con el tema propuesto es decir que su contenido es diferente al tema.

Ahora bien, los estos estudios analizados también fueron clasificados de acuerdo a la categoría de las revistas en que fueron presentados, FECYT (2017) comentan en su trabajo las clasificaciones de las revistas y su contenido; los resultados de la clasificación se resumen en Tabla 10**.**

Tabla 10

Resumen de ranking de revistas.

Clasificación	No. de Estudios	Porcentaje
Q1	12	28%
Q2	3	7%
Q3	2	5%
Q4	1	2%
N/D	25	58%
Total	43	100%

Nota. Resume cuantos artículos se asocian a revistas

Es importante indicar que existió un 58% de los estudios que no fueron publicados en revistas.

2.4 Análisis de la información

Es la actividad de lectura realizada a cada estudio seleccionado, este análisis debe extraer la información importante como por ejemplo las técnicas de minería de datos u otras utilizadas para la identificación del t*rolling*. El criterio del investigador es importante para poder resumir y sacar los aspectos clave de cada estudio (Tebes et al., 2019) para ello se requieren contestar las siguientes preguntas mostradas en Tabla 11.

Tabla 11

Preguntas de investigación complementarias

	Preguntas
1	¿Cuál es la fuente de información que se usa en los estudios para los análisis?
2	¿Qué algoritmos son los más utilizados en la detección del *trolling*?
3	¿Cuáles fueron las fuentes de datos utilizadas para los estudios?

Nota. Preguntas necesarias para identificar fuentes de datos y algoritmos

Para plasmar los datos e información relevante fue necesario construir una matriz que guarde las principales características que entregan los documentos analizados, por consiguiente, se estableció una matriz con columnas como el título, autor, año, técnicas utilizadas, DOI entre otros, necesarias para poder crear una vista de la información y sus resultados. Los atributos de esta matriz se los menciona en la Tabla 12.

Tabla 12

Matriz de atributos.

Atributos de la tabla matriz	Descripción del atributo
Autor	Autores del estudio
Titulo	Título del documento
Año	Fecha de publicación
Origen del titulo	Clasificación inicial
DOI	Código universal para codificar un documento
Abstract en inglés	Resumen en Ingles
Palabras Claves de Autor	Palabras importantes que el autor describe
Técnicas utilizadas	Algoritmos, modelos, utilizados
Detalles de las Técnicas	Resumen de la aplicabilidad de los modelos
Palabras encontradas en el documento	Palabras importantes encontradas en la búsqueda de los documentos
Abstract en español	Resumen en español
Puntos clave	Resumen de párrafos internos que destacan
Palabras Clave	Resumen de palabras importantes señaladas en el documento
Tipo	Clasificador de validez: determina si el documento es utilizado o no para el desarrollo del estado del arte
Tipo de Documento	Identifica si el documento es parte de una revista, una conferencia o un artículo
Analizado	Estado del análisis
Origen	Fuente o repositorio origen

Nota. Atributos utilizados en la matriz de relevamiento de estudios.

2.5 Hallazgos

Los hallazgos son el resultado de la investigación realizada, con ellos se puede comprender y tomar decisiones posteriores.

Esta es la actividad final en la que se desarrolla el informe del análisis sistemático, aquí se muestra claramente los resultados de los de los estudios, sus objetivos, problemas, métodos, fuentes de información utilizadas además de otros aspectos claves que han sido determinantes y podrán ser utilizados posteriormente con otros investigadores del tema.

Para una mejor organización de los hallazgos se realiza esta lista de puntos que se debe responder por cada estudio analizado, así se podrá obtener información ordenada y argumentada:

1. Una clasificación del artículo
2. Un resumen, colocando los puntos relevantes
    2.1 Autores
    2.2 Años del estudio
    2.3 Fuentes de información utilizadas
    2.4 Modelos usados
    2.5 Algoritmos utilizados
    2.6 Resultados obtenidos

Por último, estos hallazgos son validados nuevamente con el fin de garantizar que sean lo suficientemente informativos y que cumplan así los objetivos de la investigación.

CAPÍTULO 3

RESULTADOS

En este capítulo se expone en resumen cada uno de los artículos analizados y los resultados o hallazgos que se han obtenido en relación con los modelos y algoritmos utilizados para la detección del *trolling* y el ciberacoso. En cada análisis se realiza una referencia de su autor, el año de publicación del estudio, la fuente de información, métodos y técnicas utilizadas, así como los resultados obtenidos.

3.1 Análisis de los modelos utilizados para la detección del *trolling*

El *trolling* o trolear no es un comportamiento que ha aparecido en los últimos años, existen estudios que hablan de la detección de comportamientos emocionales mediante algoritmos de clasificación como lo menciona C. M. Lee et al. (2001), quienes detallan en su trabajo el haber realizado un taller en el cual se desarrollan modelos de reconocimiento de emociones negativas en la escritura, permitiendo así conocer formalmente técnicas que nos ayuden a detectar estos problemas .

Posteriormente Ververidis et al. (2004) proponen el diseño de una herramienta que pueda ser usada en la ciencia de la psicología que entre algunas cosas pueda clasificar y detectar patrones de ira, felicidad, neutralidad o tristeza. Esta herramienta usa técnicas de clasificación mediante el uso de algoritmos como redes bayesianas, adicionalmente se expone en este estudio que no se obtuvieron valores de precisión del 51.6 % los cuales no son significativos para llegar obtener a un grado de certeza sobre el reconocimiento de las emociones.

El estado emocional también juega un papel importante en el *trolling* pues por un lado están los *trolls* que por su naturaleza causan caos, pero por otro lado están los usuarios normales que también desencadenan este comportamiento al ser alterado probablemente su estado de ánimo o emocionalidad por los comentarios *trolls* es así que Cheng et al. (2017) plantean en su artículo que los usuarios en redes sociales pueden ser fácilmente influenciados, por lo tanto, surge la pregunta de que si el comportamiento del troleo es innato o situacional; en este sentido el autor señala la necesidad de crear un experimento controlado y un análisis longitudinal de gran escala para detectar este tipo de comportamiento.

Ya en el desarrollo del estudio Cheng et al. (2017) proponen una encuesta que consiste en realizar un cuestionario y un foro de discusión, el cual fue llevado a cabo en Amazon Mechanical Turk. El alcance de esta investigación estaba restringida solamente a los Estados Unidos. El resultado concluye que el *trolling* surge desde los factores innatos y situacionales, esto significa que el estado de ánimo afecta al comportamiento del troleo lo que sugiere que se mejoren los factores situacionales que conducen a este comportamiento.

En la década anterior (2010-2020) los países del Reino Unido se preparaban en su legislación con leyes que permitan contrarrestar los efectos de este comportamiento intrusivo y causante de caos, es así que Bishop (2013) publica un artículo en el que se hace referencia a los cambios que se ejecutaron en la legislación británica, estos cambios castigan los delitos llamados "*flaming trolling*"; en este trabajo no solamente se dictan leyes, sino que también existe una tabla de magnitud del *trolling* en forma de acoso como se muestra en la Tabla 13. Dotando de herramientas a la policía o a los jueces que sean parte del juzgamiento de estos delitos.

Tabla 13

Escala de magnitudes

Type	Potential Gravity	Severity	CPS score	Appropriate legal provision
TMI – Playtime	In the moment and quickly regret (finem facere)	Minor (TM 1.00-1.49)	4	Fixed penalty notice of €75
Cyber-bantering (Cyber-trolling)	Grossly Offensive, Indecent, Obscene or Menacing	Major (TM 1.50-1.99) Minor (2.00-2.49)	4 1	Fixed Penalty Notice of €150 Common law detention for breach of the peace as permitted by s.40(1) Of the Public Order Act 1986
TM2 Tactical Cyber-trickery (Cyber-trolling)	In the moment but do not regret and continue (animus restituendi) Harassment, Alarm or Distress	Major (TM2.50-2.99) Minor (TM3.O-3.49)	2 2	ASBO under s.1 the Crime and Disorder Act 1998. Harassment warning under s.4 or s.4A Of the Protection from Harassment Act 1997 or caution under relevant legislation.

Type	Potential Gravity	Severity	CPS score	Appropriate legal provision
TM3 — Strategic Cyber-bullying (Cyber-stalking)	GO out of way to cause problems, but without a sustained and planned long-tenn campaign. Harassment, Alarm or Distress (ad vita aut ademptio).	Major (TM3.50-3.99)	2	Custodial sentence under s. 127 Of the Communications Act 2003 or s. I Of the Malicious Communications Act 1988, Of 26 to 52 days or up to 18 weeks.
TM4 — Domination Cyber-hickery (Cyber-stalking)	Goes out of the way to create rich media to target one or more specific individuals Grossly Offensive, Indecent, Obscene or Menacing (ad vita aut ademptio)	Minor (TM4.0-4.49) Major (TM4.50-4.99)	2 3	Restraining order under s.5 Of the Protection from Harassment Act 1997 or related court orders, such as under The Family Law Act 1996. Custodial sentence under s. 127 of the Communications Act 2003 or s. I Of the Malicious Communications Act 1988, for up to 6 months for each act. Related custodial sentences for breaching a court order.

Nota. Escala de magnitud que determina el nivel del *trolling.* Fuente Bishop (2013).

Bastos et al. (2018) inicia su análisis contextualizando el Brexit como una palabra de tipo acrónimo para expresar Britain y Exit la simplifica el concepto de la salida de Gran Bretaña de la Unión Europea. Este tema polémico fue abordado inicialmente en junio del 2016 y tras algunas aprobaciones se ejecuta la salida en marzo del 2017. Al ser un tema polémico y analizado en un referéndum miles de personas comenzaron a publicar contenido en redes sociales alusivo al tema por lo cual se desarrolla una verdadera guerra política. En este estudio se analizaron 13.493 cuentas en Twitter en las que hubo incluso cuentas similares *bots* inmersas en los comentarios previos a las votaciones del Brexit, tratando de sesgar el voto para cualquiera de las dos partes.

En otro trabajo realizado por Bastos et al. (2018) cuestionan seriamente lo difícil que en ese año resultaba obtener información de las redes sociales para poder analizar este tipo de comportamientos, y también el obtener interfaces de conexión a las mismas. Los autores señalan que el obtener información para el estudio fue un reto significativamente alto y cuestionable hacia las corporaciones que rigen el acceso a la información pública, entonces

surge una pregunta ¿por qué las organizaciones proveedoras de los servicios de redes sociales no pueden entregar datos que ayuden a los investigadores a mitigar las guerras de desinformación?, esta pregunta no solo se presenta en este trabajo, sino también en otros estudios relacionados al tema.

También surge otra pregunta ¿por qué las técnicas de detección de información de tipo *trolling* no son tan precisas o han evolucionado lentamente?, la respuesta está quizá en el acceso y la obtención de información para estudios, por lo tanto, la guerra de desinformación *trolling* continúa creciendo (Bastos et al., 2018).

Aparecen en este escenario los denominados *bots*, estos son utilizados para transmitir contenido en forma programada, los *bots* fueron usados principalmente por las cadenas de noticias y sistemas de alerta para emitir mensajes al instante, o también cuentas que al parecer fueron gestionadas por entes como el Vaticano y que ofrecieron reflexiones del catolicismo (Bastos et al., 2018).

El trabajo realizado muestra que los *bots* no son sistemas complejos más bien son simples rutinas, sin embargo, la detección de estos todavía sigue siendo un reto para los investigadores que usan algoritmos pues todavía no ofrecen un nivel de certeza que garantice su identificación.

El estudio del *trolling* aborda una conceptualización de los *trolls* como entes que causan caos, acoso, confusión, ira; estos comportamientos han desencadenado problemas mayores como la ansiedad y la depresión, ¿pero nos preguntamos qué pasa con las víctimas del *trolling*? Tian et al. (2019) enfocan esta realidad en su estudio, tratando de determinar el impacto negativo que esto genera; se muestra que las víctimas de *trolling* estuvieron abrumadas por el troleo intenso al que estaban expuestas.

El equipo de estudio realiza 19 entrevistas a personas que tuvieron alguna experiencia de troleo en China, 17 personas fueron víctimas y las dos restantes fueron testigos, la mayoría de las víctimas describieron el tener experiencias estresantes, deprimentes o que les generó tristeza, estas experiencias fueron identificadas en sitios parecidos a Facebook o Twitter.

Las victimas tuvieron diferentes percepciones sobre las afectaciones y pudieron describir los entornos en donde hubo más puntos de acoso, el medio por el cual existe más troleo negativo en algo como la escala de intensidades afectivas que brindan las redes sociales.

Desde el punto de vista de este análisis consideramos que la muestra fue muy corta como para tener una apreciación de la afectación del *trolling* en las personas, se sugiere ampliar las muestra y las condiciones de entornos para así poder determinar quizá el efecto real.

Se ha encontrado en este análisis varios artículos científicos que hablan de las formas de conocimiento y mitigación de *trolling*, autores como Dlala et al. (2014) han realizado estudios formales que ayudan a detectar estos comportamientos mediante la utilización de algoritmos como *k-means.* En el mismo año Galán-García et al. (2014) ya describen en su estudio la necesidad de detectar perfiles falsos que son creados en las redes sociales con la finalidad de causar un daño importante en estas personas, para este estudio toman información de redes como Twitter, (Herrero et al., 2014).

Los tentáculos del *trolling* se extienden también en los centros educativos en donde los estudiantes son víctimas del acoso que en algunas ocasiones pueden detonar en suicidios o afectaciones psicológicas, por tanto, Herrero et al. (2014) proponen también abordar un análisis de este comportamiento en centros educativos. Es así que se realiza un análisis de detección del *trolling* en Springer International Publishing Suiza (2014), en este trabajo ya se utilizan varios algoritmos de regresión lineal y de clasificación para evaluar cuál de ellos podría clasificar mejor los *Trolls; y,* para el desarrollo de este caso se obtuvo información de Twitter que sirvió para el análisis y la obtención de niveles importantes de clasificación.

Pues no solamente en redes sociales se puede encontrar a los *Trolls* que estén insertando comentarios fuera de sitio, también se ha encontrado este tipo de comportamiento en sitios web donde existe la posibilidad de tener grupos de discusión, es así que De-La-Pena-Sordo et al. (2015) proponen que los sistemas informáticos que permiten los foros de discusión tengan un filtro de contenido anti *trolling*, para este objetivo se usan modelos que utilizan algoritmos como KNN, Naïve Bayes, SVM, J48 y Random Forest, estos son usados para determinar el contenido malicioso.

El *trolling* en línea ha tomado algunas formas, pero en esencia son publicaciones acosadoras, ofensivas, amenazantes e intimidantes, en este sentido el centro de investigación PEW descubre que a partir del 2013 existen algunas estadísticas importantes.

- El 73% de las personas han sido testigos de un tipo de acoso.
- El 40% lo han experimentado directamente.

Y de estos usuarios se presentan las siguientes estadísticas más alarmantes.

- El 53% a distinguido intentos para avergonzar a propósito.
- El 25% ha identificado amenazas físicas.
- El 24% ha percibido un acoso que se ha mantenido en el tiempo.
- Y el 19% a distinguido acoso sexual a otras personas.

Todos estos resultados fueron realizados por Golbeck et al. (2017).

A más de los filtros que se han identificado anteriormente que servirían para identificar *trolls* se analiza la posibilidad de detectar vulnerabilidades antes que los *trolls* actúen en un grupo específico donde se trata un tema de discusión que es vulnerable, este estudio lo realiza Tsantarliotis et al. (2016) utilizando datos de publicaciones de la red social Reddit; el objetivo es recordar las publicaciones que fueron vulnerables al *trolling*.

Otro importante trabajo realizado por Mutlu et al. (2016) en el que además de proponer un análisis de *trolls* en redes como Twitter en el cual se propone la identificación de la llamada conciencia de terror, el desarrollo del estudio se basa en análisis a las discusiones que se abordan en todas las partes del mundo relacionadas al terrorismo. Esta medida de conciencia determina que tanto conocen las personas sobre las acciones terroristas. La investigación se realiza en Turquía debido a que es uno de los países que sensibiliza a su población en estos temas, y el trabajo de investigación concluye en un importante rendimiento del modelo utilizando el algoritmo C4.5 de árboles de decisión.

Si bien es cierto la afectación del *trolling* hoy en día es global, existe un segmento en el cual se ha desarrollado mayormente esta práctica, este segmento es el político en el cual encontramos activistas políticos radicales que tratan de desacreditar las acciones, pensamientos

o propuestas de los diferentes actores políticos; debido a este comportamiento Chang et al. (2016) plantean un método de clasificación no supervisado que toma características estructurales y de línea de tiempo para la detección de *trolls.*

El estudio se centra no solamente en la detección de rumores, sino en los usuarios que los generan, a estos lo llaman los autores "usuarios extremos"; como se había mencionado anteriormente existen características únicas que comparten estos usuarios como lo son la cantidad de seguidores, una alta frecuencia de tuiteo y mostrar mucha pasión en temas determinados. El análisis para para encontrar ese tipo de usuarios usa la clusterización basada en reglas y entrega valores de reconocimiento bastante significativos a la hora de clasificar. A continuación, se muestra la Tabla 14 que resume el modelo de clasificación enfocado a los tuis generados por Clinton y Obama.

Tabla 14

Clasificación de resultados

The Clinton dataset					
Rule	Detected	TP	Precision	Recall	F1Score
Rule 1	20	13	0.65	0.76	0.70
Rule 2	24	15	0.63	0.88	0.73
Rule 3	7	7	**1.00**	0.41	0.58
Rule 4	21	14	0.67	0.82	0.74
Rule 5	22	15	0.68	0.88	**0.77**

The Obama dataset					
Rule	Detected	TP	Precision	Recall	F 1Score
Rule 1	13	11	0.85	0.85	0.85
Rule 2	11	10	0.91	0.77	0.83
Rule 3	5	5	1.00	0.38	0.56
Rule 4	9	9	1.00	0.69	0.82
Rule 5	10	9	0.90	0.69	0.78

Nota. Clasificación de los resultados de las reglas con mejor desempeño en la detección de rumores de noticias. Fuente Chang et al. (2016).

Las noticias se generan a diario y son mostradas por los diferentes medios de comunicación o empresas que se dedican a exponer noticias, muchas de ellas generan un gran contenido de

mensajes de opinión de sus lectores en ellas se arman verdaderos foros de discusión y de las cuales se han identificado que existen algunos usuarios que con el simple hecho de conocer la forma de opinión en sus mensajes se podría intuir si es un *troll* o no. Por este motivo Authors (2017) señala en su estudio el éxito que tuvo en la detección y clasificación de *trolls*. Su trabajo utilizó algoritmos de RL y SVM y los autores califican como exitoso el resultado, obteniendo valores significativos para la detección del *trolling.* Se menciona también que en todo este desarrollo se utilizan datos estáticos de una fuente no determinada, los resultados obtenidos se muestran en la Tabla 15.

Tabla 15

Resultados de clasificación

Systems	Precision	Recall	F_1	Accuracy
All features	83.42	78.00	80.40 (+30.40)	81.31 (+31.31)
Bag of words	76.40	74.31	75.27 (+25.27)	75.62 (+25.62)
Metadata	70.34	77.85	73.79 (+23.79)	72.54 (+22.54)
Baseline	50.00	50.00	50.00	50.00

Tabla 18. Resultados de clasificación de *trolls* vs no *trolls.* Fuente Authors (2017).

El trabajo concluye que a futuro se deberá de replicar este análisis, pero con datos que tengan mayor variabilidad, entro idioma y con la incorporación de redes neuronales convolucionales.

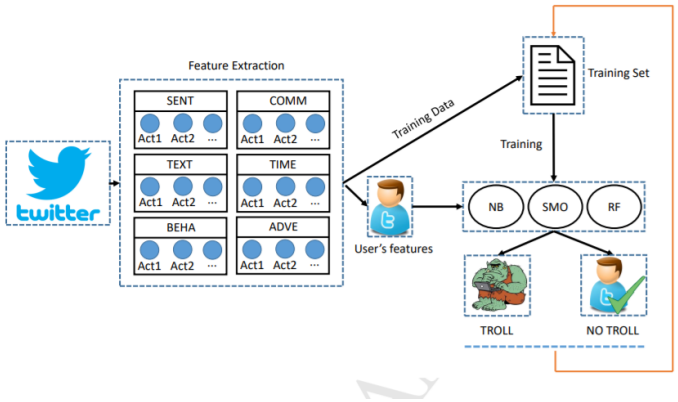
Fornacciari et al. (2018) proponen en su artículo la creación de un sistema llamado Trollpacifier que es un sistema holístico para detectar *Trolls*, lo que significa que su análisis se basa desde el punto de vista de las múltiples interacciones que lo caracterizan. El autor ha identificado seis características basadas en el estilo de la escritura:

- El sentimiento (sent)
- El comportamiento de usuario (beha)
- Las interacciones sociales (comm)
- Publicidad de contenido externo (adve)
- El tiempo de publicación (time)
- El estilo de la escritura (text)

La arquitectura y el esquema para el análisis se muestra de la siguiente manera representada en la Figura 2**.**

Figura 2

Arquitectura para el análisis

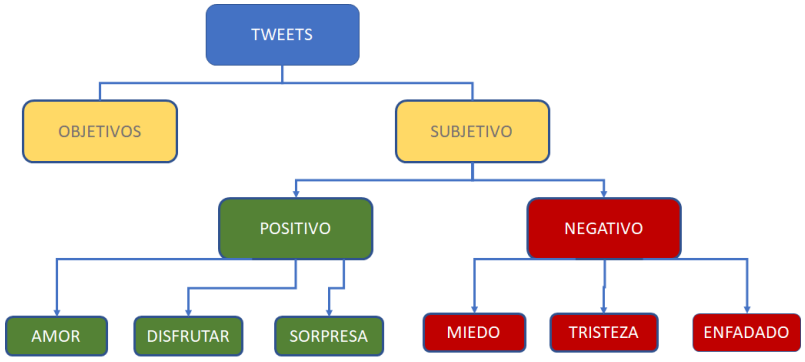


Nota. Representa la arquitectura de fuentes de datos y el análisis con algoritmos. Fuente Fornacciari et al. (2018).

Se muestra adicionalmente la clasificación de las emociones graficada en la Figura 3

Figura 3

Clasificación emocional



Nota. Clasificación de los tuits relacionados con las emociones. Fuente Fornacciari et al. (2018).

Las respuestas en los seis grupos analizados son alentadoras bajo tres algoritmos analizados como se muestra en la Tabla 16**.**

Tabla 16

Resultados de los algoritmos SMO, NB y RF

	Accuracy %			F- measure			Kappa			AUC			Recall		
	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF
SENT	78.80	72.31	78.97	0.79	0.71	0.79	0.58	0.45	0.58	0.79	0.79	0.87	0.77	0.67	0.79
TIME	67.84	61.28	75.65	0.57	0.41	0.75	0.36	0.23	0.51	0.68	0.77	0.83	0.44	0.27	0.74
TEXT	67.56	64.97	68.47	0.66	0.63	0.69	0.35	0.30	0.37	0.68	0.69	0.74	0.64	0.59	0.70
BEHA	75.88	58.62	79.59	0.75	0.70	0.80	0.52	0.17	0.59	0.76	0.82	0.87	0.70	0.97	0.79
COMM	80.45	74.96	83.16	0.80	0.72	0.83	0.61	0.50	0.66	0.80	0.83	0.91	0.78	0.64	0.83
ADVE	78.07	71.70	85.01	0.78	0.67	0.85	0.56	0.43	0.70	0.78	0.79	0.92	0.76	0.59	0.85
TOT	95.52	80.25	88.28	0.95	0.78	0.88	0.91	0.60	0.77	0.96	0.90	0.96	0.95	0.69	0.89

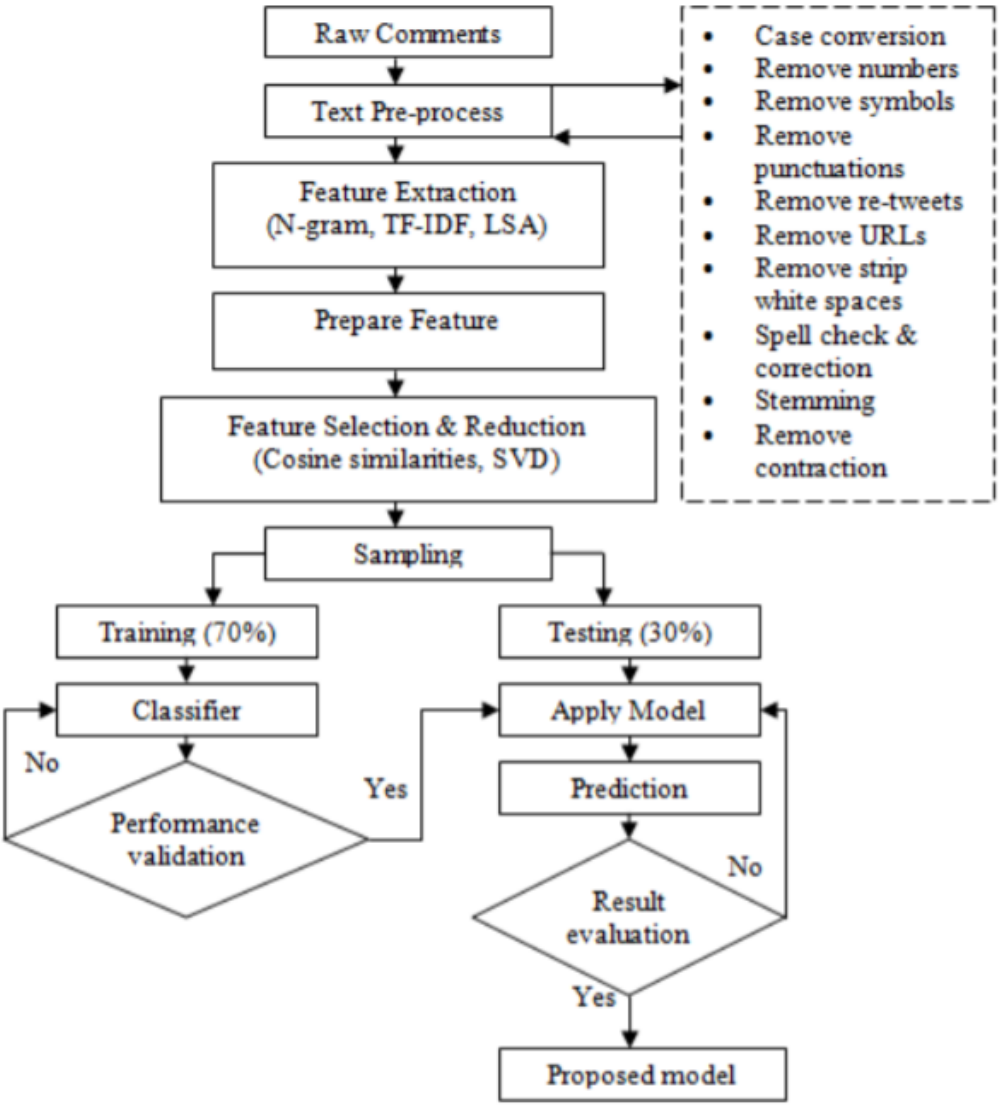
Nota. Resultados de los diferentes algoritmos utilizados en el análisis. Fuente Fornacciari et al. (2018).

Otro estudio que ayuda a clasificar el contenido de odio o racismo e injuria es el desarrollado por Rasel et al. (2018) quienes realizan un modelo de clasificación que se enfoca en dividir a los comentarios de redes sociales en tres grupos: ofensivo, odio y ninguno de ellos, el valor obtenido en la precisión sobre la clasificación fue del 93%.

Para que este modelo tenga el resultado esperado el autor indica la importancia del entrenamiento del modelo para obtener valores adecuados de respuesta, se propone un flujo de actividades para el análisis como lo muestra la Figura 4.

Figura 4

Flujo de actividades

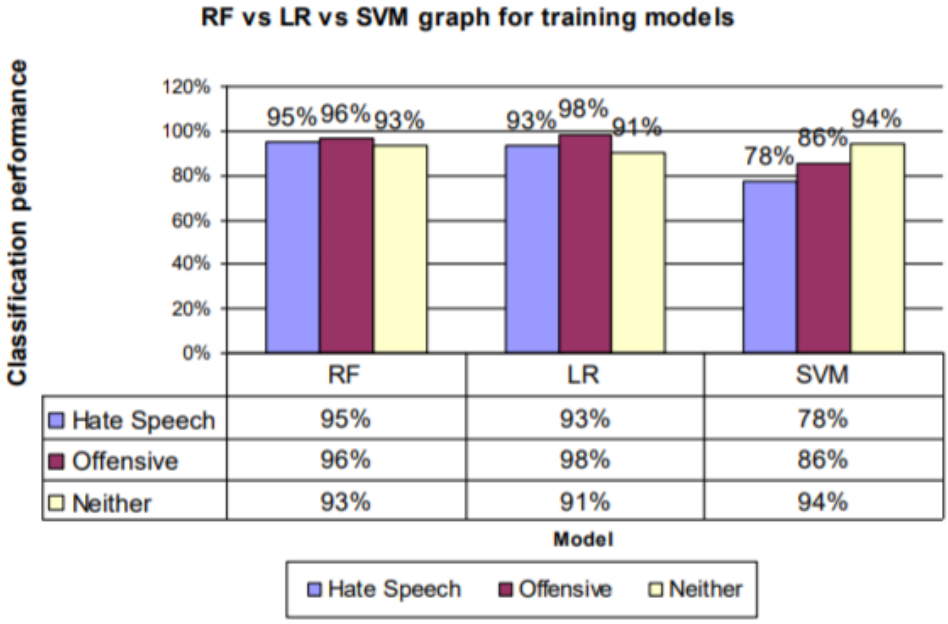


Nota. Flujo de propuesta para la detección de contenido *trolling*. Fuente Rasel et al. (2018).

El modelo al aplicar los algoritmos seleccionados entrega un porcentaje de precisión en la clasificación significativo, superior en algunos casos al 93% como lo muestra la Figura 5.

Figura 5

Cuadro de resultados



Nota. Resultados de los algoritmos utilizados RF, LR y SVM. Fuente Rasel et al. (2018).

Bellan et al. (2018) analizan la posibilidad de detectar comentarios inapropiados en periódicos profesionales, en su estudio recogen datos de comentarios inapropiados sobre los cuales se prueban tres modelos clásicos de aprendizaje automático, para ello utilizan una técnica llamada distorsión de texto la misma que mejoró significativamente la clasificación entregando información valiosa como las características basadas en el léxico.

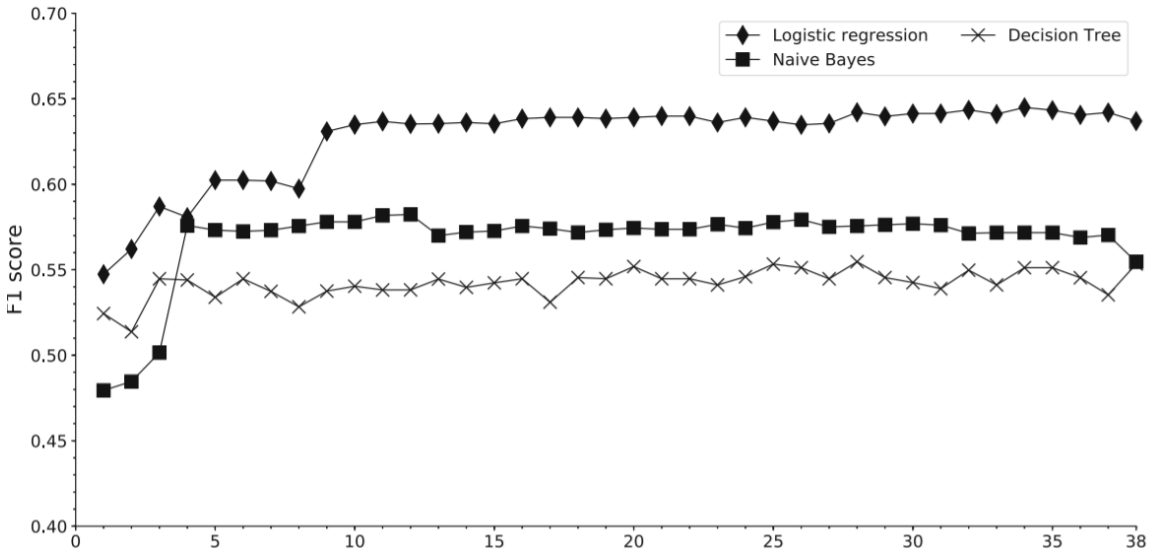
El periódico The Guardian fue el elegido para obtener datos desde noviembre del 2017 hasta enero del 2018, estos datos eran recopilados de las secciones que estaban disponibles al público, aunque los administradores del sitio web borran contenido inapropiado sigue habiendo mucha información. Para obtener información confiable fue necesario hacer un seguimiento diario a los datos y clasificar este contenido en dos categorías: comentarios apropiados vs comentarios inapropiados, adicionalmente para probar otro contexto de referencia se escogió a Twitter de donde también se obtuvo datos.

Ya en el desarrollo del trabajo, se utilizaron tres algoritmos de aprendizaje automático, Naïve Bayes, Regresión Logística y Arboles de Decisión.

El análisis muestra que la Regresión Logística entrega un buen nivel de rendimiento sobre el conjunto de datos asignado, cerca del 65% como lo muestra la Figura 6, mientras que la Tabla 17 muestra las puntuaciones de cometarios y tuits y las puntuaciones más altas de cada algoritmo marcadas con negrita.

Figura 6

Resultados obtenidos en entrenamiento



Nota. *Score* F1 y características del entrenamiento. Fuente Bellan et al. (2018).

Tabla 17

Resultados de los algoritmos

Algorithm	Sets	(k)	F1 comments	F1 tweets
Logistic regression	1	1000	.547	.397
	2	3500	.562	.333
	3	5000	.587	.341
	8	3500	.597	**.509**
	10	1000	.635	.499
	34	3500	**.645**	.491
Decision tree	1	3500	.524	.333

Algorithm	Sets	(k)	F1 comments	F1 tweets
	2	3500	.514	.333
	3	5000	.545	.333
	26	5000	.551	.489
	29	3500	.555	.333
Naïve Bayes	1	3500	.479	.333
	'2	5000	.485	.333
	3	5000	.501	.398
	4	3500	.578	.333

Nota. Resultados de los algoritmos utilizados. Fuente Bellan et al. (2018).

Ahora bien, es importante mencionar que los tentáculos de los *Trolls* se extienden desde una red social hasta sistemas de ayuda a personas, uno de los más importantes ambientes que un *Troll* persigue es el político, y quizá los intereses políticos creen esta necesidad de *trolling*.

El desvirtuar a candidatos políticos y sesgar comportamientos de votantes es algo común en los *trolls* un estudio realizado por Kellner et al. (2019) muestran una posible injerencia de los países en las elecciones presidenciales de Estados Unidos en el 2016; y, que también han actuado en las elecciones federales de Alemania. Se analizaron campañas destinadas para afectar la opinión pública, en el artículo se comenta el haber encontrado 79 *Trolls* en esta campaña, se usó para ello un detector de comportamientos que permite encontrar cuentas sospechosas.

Para el desarrollo del análisis Kellner et al. utilizaron los datos de tuits generados entre enero y septiembre del 2017, son aproximadamente 9,5 millones de tuits de los cuales se obtiene una muestra significativa, estos datos fueron usados para comenzar el análisis aplicado a las elecciones federales de Alemania.

En base a los hallazgos se realiza la identificación del comportamiento automático de la cuenta responsable de amplificar los debates políticos, para esto se aplica un aprendizaje supervisado que ayuda a encontrar esos *bots*, siendo esto un reto que se mantiene hasta la actualidad.

Los resultados se obtuvieron del análisis con varios modelos de clasificación como se expone en la Tabla 18.

Tabla 18

Resultados obtenidos

Classifier	Avg. FI-Score	Avg. AUC
Gradient Boosting	0.891 ± 0.033	0.976 ± 0.011
Random Forest	0.861 ± 0.039	0.972 ± 0.013
AdaBoost	0.885 ± 0.035	0.971 ± 0.013
SVC	0.851 ± 0.038	0.949 ± 0.021
Logistic Regression	0.841 ± 0.043	0.946 ± 0.021
Linear SVC	0.821 ± 0.040	0.934 ± 0.025
KNeighbors	0.704 ± 0.060	0.871 ± 0.034

Nota. Resultados de los algoritmos utilizados para el análisis. Fuente Kellner et al. (2019).

El tema de discusión se centra no solo en la detección de *Trolls* sino también en el sesgo que sus publicaciones (*trolling*) interfieren en las decisiones humanas, por consiguiente, todos los esfuerzos que se hagan para contrarrestar esta realidad son necesarios y deben seguir incrementándose y madurando en el tiempo, sobre todo en las elecciones de dignidades políticas. En campañas electorales que han usado redes sociales se ha identificado que todas las armas para ganar una elección son valiosas para quienes en realidad quieren tener un espacio de poder sin importar las consecuencias.

Cambiando de tema, se analiza ahora el trabajo de Marouf et al. (2019) quienes identifican las redes como medios beneficiosos y al mismo tiempo perjudiciales por la influencia de los comentarios de los *trolls* en el cual se visualiza la detección de *trolls* desde otro punto de vista. El objetivo de este trabajo es identificar al perfil detrás del *troll*, esta realidad es ejemplificada con algunas celebridades quienes son objeto constante de *trolling* que en algunos casos les han causado hasta problemas médicos por los rumores o difamaciones inventadas en estos sitios.

El enfoque es detectar los rasgos de personalidad de los *trolls* basados en estudios anteriores que utilizaron herramientas de software psicolingüístico las cuales han logrado obtener buenos resultados. Estas herramientas contienen un conjunto de datos o palabras (sinónimos y antónimos en diferentes categorías psicológicas); *Linguistic Inquiry y Word Count* (LIWC) son

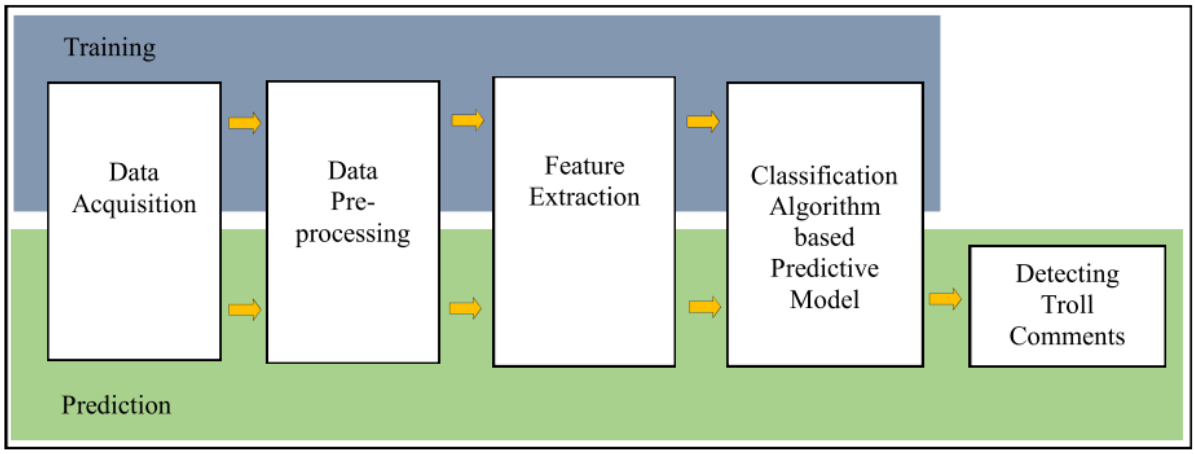
útiles en un contexto de predicción de rasgos de personalidad, también existen otras herramientas como SlangNet.

El alcance del trabajo fue identificar comentarios *trolls* publicados para desprestigiar personajes importantes, para estos análisis se usaron las herramientas psicolingüísticas mencionadas anteriormente en conjunto de un modelo de aprendizaje supervisado.

Los autores sugieren la utilización del siguiente marco metodológico para desarrollar el proceso de análisis como se muestra en la Figura 7.

Figura 7

Proceso de análisis



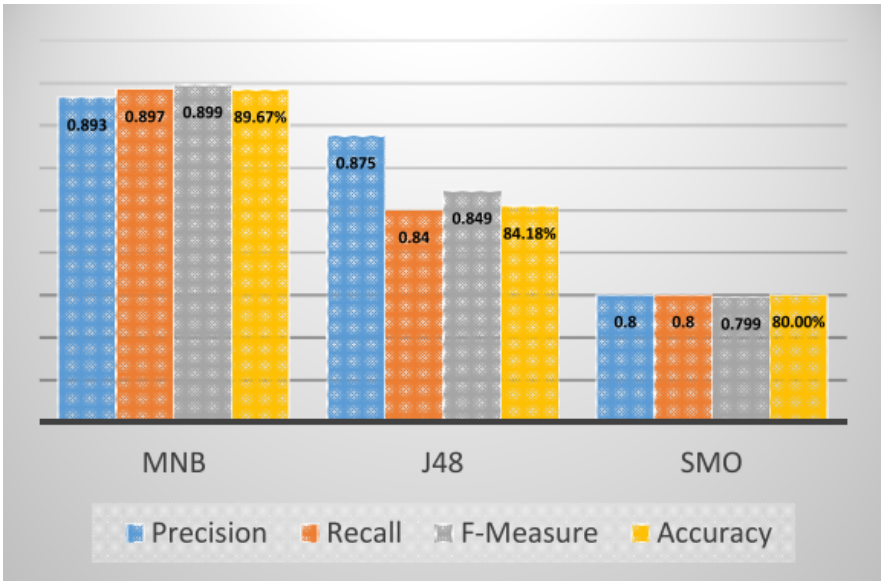
Nota. Gráfico que muestra el proceso para realizar el análisis. Fuente Marouf et al. (2019).

La recopilación de datos se realiza mediante la aplicación Graph API para obtener los estados de "Me gusta" y los comentarios de Facebook de los usuarios que estén disponibles, se lograron extraer 7000 cometarios de 140 perfiles.

Para el análisis se utilizaron los modelos MNB, Árbol de Decisión (J48) y Support Vector Machine, adicionalmente se utiliza la herramienta Weka para construir los modelos de predicción, los rendimientos obtenidos son expuestos en la Figura 8.

Figura 8

Resultados obtenidos



Nota. Resultado de los algoritmos utilizados MNB, J48 y SMO. Fuente Marouf et al. (2019).

El resultado muestra un rendimiento mejor utilizando el algoritmo de MNB logrando una precisión del 89.6%, un porcentaje significativamente bueno para la clasificación de *trolls*.

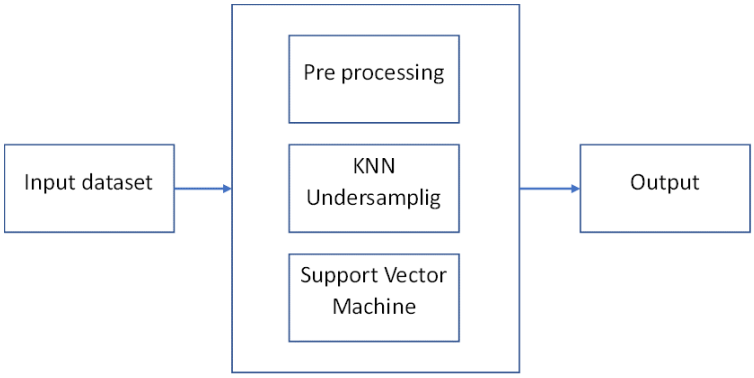
En años recientes Luqyana et al. (2019) al igual que otros investigadores buscan clasificar adecuadamente los tuis normales de los tuis de tipo *trolling*, los autores consideran que es un reto complejo ya que las diferencias de lenguaje podría ser un problema.

Algo importante que señalan en su estudio son las consecuencias del *trolling*, pues solamente en el Reino Unido se encuentran casos de acosos a familias quienes tuvieron la fortaleza de realizar acusaciones directas hoy sus acosadores están detenidos.

Respecto a los análisis para determinar los tuis de tipo *trolling* existieron algunos problemas como lo son los datos desequilibrados, ya que estos pueden incidir en el resultado, por lo tanto, sugieren utilizar el modelo KNN-undersampling que en estudios previos ha dado mejores resultados para lograr un equilibrio en la data; ya en el análisis se usa SVM para la clusterización como se presenta en la Figura 9.

Figura 9

Marco de referencia

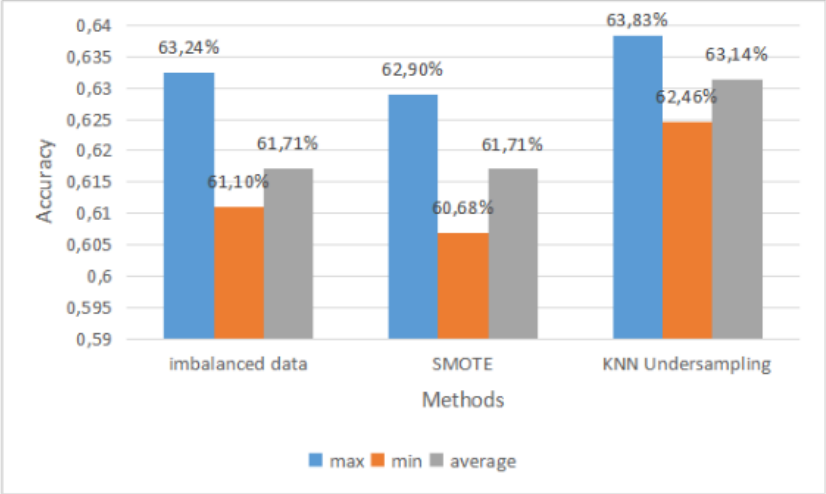


Nota. Marco de referencia utilizado para el análisis de datos. Fuente Marouf et al. (2019).

El conjunto de datos utilizado para el análisis es obtenido de la red Twitter y para la realización del experimento se usa la herramienta Python para resolver los datos desequilibrados usando KNN-Undersampling, el dataset analizado contiene 13.000 tuits positivos y 7000 tuits negativos, el resultado del balance se muestra en la Figura 10.

Figura 10

Resultados obtenidos



Nota. Resultados de los algoritmos SMOTE y KNN Undersampling utilizados en el análisis de datos. Fuente Marouf et al. (2019).

SVM clasifica los tuits de trols vs los normales mientras que KNN-Undersampling se utiliza para mejorar ligeramente el rendimiento en datasets desequilibrados; en este sentido KNN-Undersampling tiene un mejor rendimiento que SMOTE; los autores finalizan sus conclusiones indicando que en futuros estudios se utilizará redes neuronales para el procesamiento de lenguaje natural.

Ahora, continuando con la temática política Badawy et al. (2019) intentan demostrar que existió injerencia Rusa en las elecciones de los Estados Unidos en el 2016 utilizando la red social Twitter, el trabajo se centra en dos fases, la primera es probar si existen usuarios que difunden contenido *troll* y la segunda es identificar los usuarios que de forma intencional o no difunden contenido malicioso.

El estudio comienza recopilando información de publicaciones, alrededor de 43 millones de tuis entre el 6 de septiembre y el 9 de noviembre del 2016.

Se utiliza una lista de 2752 cuentas de Twitter ya identificadas como como *trolls* rusos, información que fue recopilada por el congreso de los Estados Unidos como lo presenta la Tabla 19.

Tabla 19

Datos estadísticos encontrados

	Value
# of Russian trolls	2.735
# of trolls in our data	221
# of trolls wrote original tweets	85
# of original trolls' tweets	861

Nota. Número de *trolls* rusos encontrados. Fuente Badawy et al. (2019).

Posteriormente definen las características de los usuarios para poder predecir aquellos que esparcen información, estas características se detallan en la Tabla 20.

Tabla 20

Características de perfiles

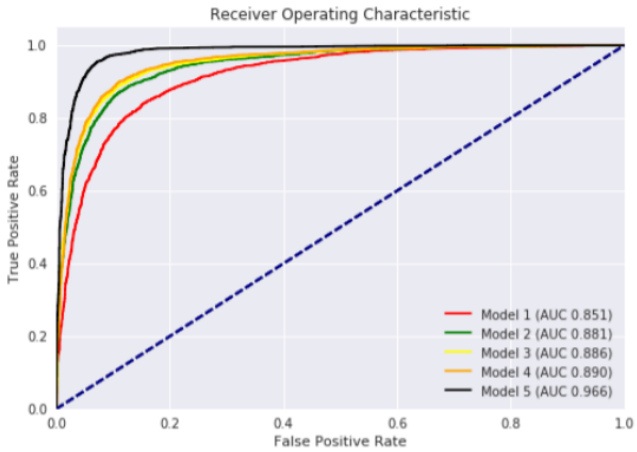
Metadata	LIWC	Engagement	Activity	Other
# of followers	Word Count	Retweet variables	#of characters	Political Ideology
# of favorites	Positive Emotion	Mention variables	#of hashtags	Bot Score
# of friends	Negative Emotion	Reply variables	#of mentions	Tweet Count
Status count	Anxiety	Quote variables	#of urls	
Listed count	Ager			
Default Profile	Sadness			
Geo-enabled	Analytic			
Background-image	Clout			
Verified	Affection			
Account Age	Tone			

Nota. Características que describen los usuarios sobre los perfiles que esparcen información. Fuente Badawy et al. (2019).

Para validar los resultados obtenidos se usó un algoritmo de propagación de etiquetas conjuntamente con una validación cruzada estratificada aplicada a un conjunto de semillas de 29k. El algoritmo fue entrenado con 4 de los 5 quintos de las semillas y en el quinto restante se usa como prueba, se concluye que los valores de precisión fueron relativamente significativos, la precisión y la recuperación fueron mayores al 91%, y para el Gradient Boosting que es un clasificador que obtuvo el 85% y 96% de rendimiento en la eliminación de bots ilustrada en la Figura 11.

Figura 11

Resultados obtenidos



Nota. Resultados obtenidos en el análisis utilizando el algoritmo Gradient Boosting. Fuente Badawy et al. (2019).

Un año después de que Badawy et al. (2019) realiza la investigación sobre la injerencia Rusa en las redes sociales Im et al. (2020) realizan un trabajo similar para demostrar participación de *Trolls* Rusos en redes sociales.

Para realizar este estudio se utiliza a la red social Twitter como fuente de datos, tomando una muestra aleatoria de 171.291 de estas cuentas como lo detalla en la Tabla 21.

Tabla 21

Resumen del cuentas y tuits

	Russian Trolls	**Control Accounts**
Total # Of Accounts	2,286 (1.4%)	171,291 (98.6%)
Total # Of Tweets	346,711	29,960,070
Avg Account Age (days)	1679.2	2734.5
Avg # of Followers	1731.8	1197
Avg # of Following	952	536.5

Nota. Datos de las cuentas y tuits. Fuente Im et al. (2020).

En el desarrollo se utilizó algoritmos de regresión logística, árboles de decisión y árboles de decisión potenciados adaptativos como se ilustra en la Tabla 22. El resultado mostró que el 2.6% de las menciones de los principales periodistas en la red, fueron ocupadas por *trolls* rusos, quizá por el corto tiempo desde este hallazgo se podría intuir que aún existen *trolls* rusos sigan activos en la actualidad.

Tabla 22

Resultados obtenidos

Algorithm	Precision	Recall	F1	AUC	Accuracy
LR	0.785	0.835	0.809	0.989	0.994
DT	0.911	0.775	0.837	0.926	0.996
ADT	0.942	0.884	0.912	0.912	0.998

Nota. Resultado de los algoritmos LR, DT y ADT. Fuente Im et al. (2020).

Dewhurst et al. (2020) realizan una investigación sobre la injerencia de las potencias extranjeras en elecciones nacionales, comentan que este tipo de interferencias son una muestra de una interacción estratégica y política de los estados unidos.

Básicamente el estudio analiza fuentes de información de la red social Twitter, que al igual que otros estudios utiliza modelos analíticos para obtener niveles de clasificación y de certeza que puedan identificar en los conjuntos de datos la potencial aparición de *trolls*, los modelos que se distinguen en el estudio son el *No-U-Turn Sampler* parecido al método HMC o MCMC, y el Gaussian Process Bayesian.

Uno de los artículos más recientes es el desarrollado por Islam et al. (2020), quienes abordan en su trabajo las técnicas que ayuden a detectar mensajes abusivos y que generan intimidación en línea mediante la fusión del procesamiento del lenguaje natural y las técnicas de aprendizaje automático.

Islam et al. (2020) comentan que las redes sociales tienen un lado beneficioso, estas permiten el desarrollo comunicacional rápido se dice que estamos a un clic de una noticia, de una oferta de trabajo, de un apoyo a una persona necesitada. Las redes utilizadas en este trabajo son: Facebook, Instagram, Twitter y TikTok. Pero, así como hay un lado positivo el lado negativo

revela el bullying y el ciberacoso que se traduce en una palabra *trolling*, países como los Estados Unidos el troleo se ha vuelto común; un dato importante indica que el 50% de los jóvenes experimentan o han experimentado un ciberacoso, estos casos podrían tener una relación de impacto físico o psicológico con la víctima, por tanto, los esfuerzos para proteger a los adolescentes deben ser significativante mayores.

Ya en el desarrollo se explica que existe una complejidad en el procesamiento del lenguaje natural por lo que proponen preparar el texto mediante dos caracteristicas:

Bag.of-work: Debido a que los algoritmos no pueden trabajar con el texto sin formato, se debe convertir el texto en vectores o números

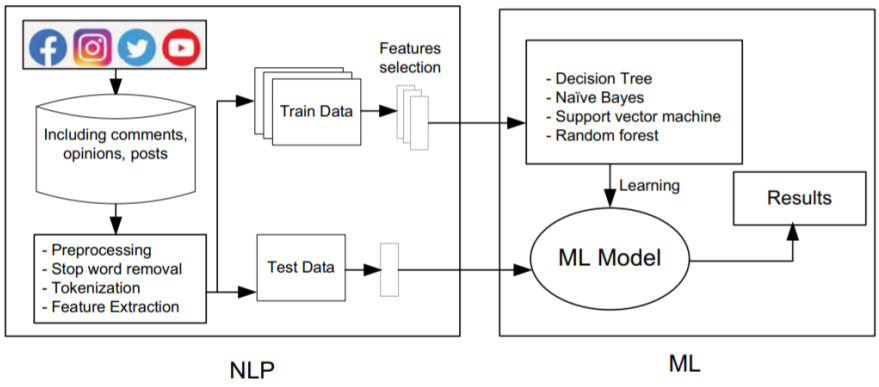
TF-IDF: Es una estadística en la que el objetivo es evaluar la importancia de una pabla en un documento, es decir que una palabra podría tener más importancia por frecuencia en que se repite.

Islam et al. (2020) utilizaron varios algoritmos de aprendizaje automático como el Decision Tree, Naïve Bayes, Random Forest, y Support Vector Machine. Los dataset utilizados para el análisis son una recopilación de comentarios de Facebook y Twitter.

Para el análisis necesitamos un esquema que nos permita obtener las clasificaciones correctas, por tanto, el autor propone lo siguiente modelo ilustrado en la Figura 12.

Figura 12

Marco de referencia para el análisis

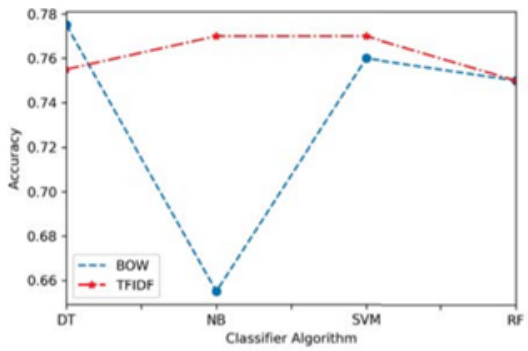


Nota. Marco de referencia propuesto para la detección del trolling. Fuente Islam et al. (2020).

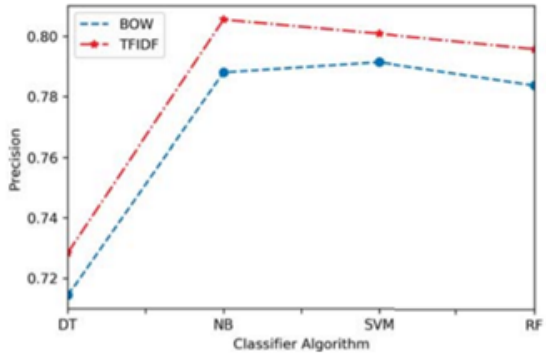
Al analizar los algoritmos se visualiza que SVM supera los demás obteniendo mejores precisiones como se muestra en la Figura 13.

Figura 13

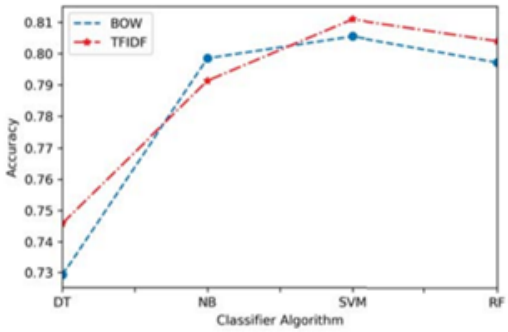
Resultados obtenidos



Exactitud de los algoritmos en el Dataset 1



precisión de los algoritmos en el Dataset 1



Exactitud de los algoritmos en el Dataset 2

Nota. Resultados de la clasificación de los algoritmos utilizados DT, NB, SVM, RF. Fuente Islam et al. (2020).

El autor concluye la necesidad de buscar automatizar estas técnicas de aprendizaje supervisado y volverlas autónomas, para así poder erradicar el ciberacoso y sus problemas sociales.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) es un marco de referencia que apoya a los grupos de investigadores a documentar de manera transparente las investigaciones que realizaron los autores y cuáles fueron los hallazgos. PRISMA permite

llevar una investigación, Sarsam et al. (2020) utilizan este framework para realizar su trabajo, en el cual se revisan los diferentes algoritmos que ayudarán a detectar el sarcasmo en Twitter.

Una extensa búsqueda condujo a los investigadores a clasificar en dos grupos los algoritmos:

- Algoritmos de aprendizaje automático avanzados denominados AMLA.
- Algoritmos de aprendizaje automático personalizados denominados CMLA.

El resultado del estudio de los algoritmos es importante, entrega una visión global de los mejores algoritmos que se adaptan al reconocimiento de sarcasmo o t*rolling.*

- SVM con mayor frecuencia = 22.58%.
- LR = 19,35%.
- Naïve Bayes= 9,67%.
- Random Forest 6,45%.

AMLA es el mejor grupo de algoritmos que entregan mayor precisión vs CML como se expone en la Tabla 23.

Tabla 23

Resultados de los Análisis a los estudios realizados

No	Autores	Enfoque	Algoritmo	Categoría
1.	Davidov et al. (2010)	Labeling based on discrete range of 1 to 5 where 5 indicates a clearly sarcastic sentence and one indicates a clear absence of sarcasm.	Semi-supervised sarcasm identification algorithm (SASI)	**CMLA**
2.	González-Ibánez et al. (2011)	Sarcastic (S), positive (P) and negative (N)	SVM	**AMLA**
3.	Riloff et al. (2013)	Polarity: positive and negative tweets	Bootstrapping	**CMLA**

No	Autores	Enfoque	Algoritmo	Categoría
4.	Kovaz et al. (2013)	Sarcastic and nonsarcastic	Logistic regression	**AMLA**
5.	Ptáček et al. (2014)	Sarcastic and nonsarcastic	Maximum Entropy (MaxEnt)	**CMLA**
6.	Tungthamthiti et al. (2014)	Sarcastic and nonsarcastic (when a tweet contains contradiction of sentiment polarity without coherence between them, it could be regarded as nonsarcastic tweet)	SVM	**AMLA**
7.	Rajadesingan et al. (2015)	Polarity: positive and negative tweets (sarcastic or nonsarcastic)	Sarcasm Classification Using a Behavioral modeling Approach SCUBA	**CMLA**
8.	Bouazizi and Ohtsuki (2015)	Polarity: positive and negative tweets	SVM	**AMLA**
9.	Joshi et al. (2015)	Sarcastic or nonsarcastic	LibSVM with RBF *kernel*	**CMLA**
10.	Ghosh et al. (2015)	Positive or negative tweets	SVM	**AMLA**
11.	Cerezo-Costas and Celix-Salgado (2015)	Polarity: positive, negative, and neutral	Logistic Regression	**AMLA**
12.	Bamman and Smith (2015)	Sarcastic and nonsarcastic	Logistic Regression	**AMLA**
13.	Barbieri et al. (2015)	Polarity: positive, negative, and neutral	SVM	**AMLA**
14.	Signhaniya et al. (2015)	Sarcastic and nonsarcastic	SVM	**AMLA**
15.	Jain and Hsu (2015)	Sarcastic and nonsarcastic	Logistic Regression	**AMLA**
16.	Bouazizi and Ohtsuki (2016)	Polarity: positive and negative tweets	Random Forest	**AMLA**
17.	Amir et al. (2016)	Sarcastic and nonsarcastic	Content and User Embedding	**CMLA**

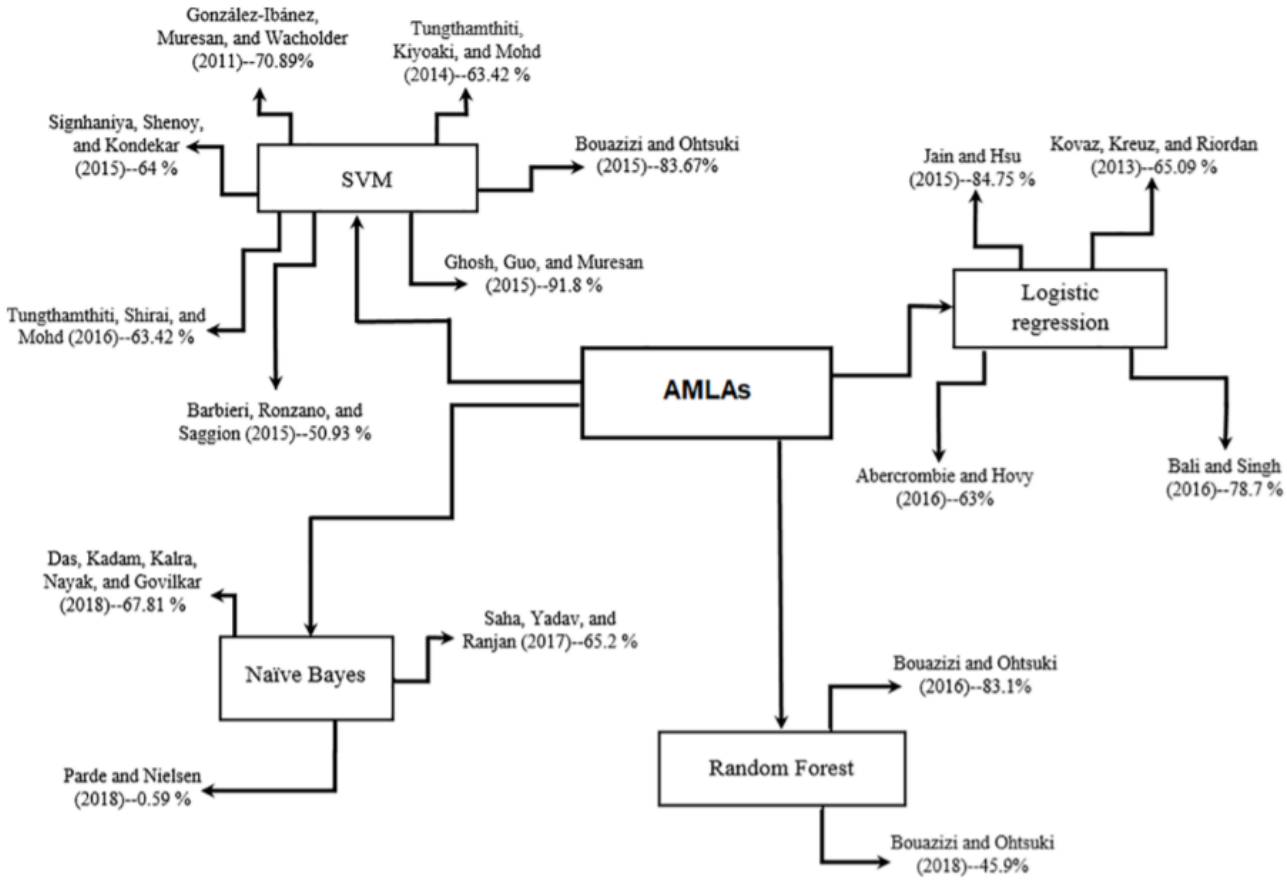
No	Autores	Enfoque	Algoritmo	Categoría
			Convolutional Neural Network CUE- CNN	
18.	Zhang et al. (2016)	Sarcastic and nonsarcastic	Gated recurrent neural network GRNN	**CMLA**
19.	Poria et al. (2016)	Sarcastic and nonsarcastic	CNN-SVM	**CMLA**
20.	Abercrombie and Hovy (2016)	Sarcastic and nonsarcastic	Logistic regression	**AMLA**
21.	Ghosh and Veale (2016)	Sarcastic and nonsarcastic	A combination of CNN; Long short-term memory LSTM network; and Deep neural network (DNN)	**CMLA**
22.	Bali and Singh (2016)	Sarcastic and nonsarcastic	logistic regression	**AMLA**
23.	Tungthamthiti et al. (2016)	Sarcastic and nonsarcastic	SVM	**AMLA**
24.	Saha et al. (2017)	Polarity: positive, negative, or neutral	Naïve Bayes	**AMLA**
25.	Prasad et al. (2017)	Sarcastic or nonsarcastic	Gradient Boost	**CMLA**
26.	Tay et al. (2018)	True and false	Multi-dimensional Intra-Attention Recurrent Network MIARN	**CMLA**
27.	Ren et al. (2018)	Polarity: positive and negative tweets	MODEL-KEY	**CMLA**
28.	Parmar et al. (2018)	Polarity: positive and negative tweets	Feature-based Composite Approach (FBCA)	**CMLA**
29.	Das et al. (2018)	Labels are two types 0 and 1 indicating the sentence being not sarcastic and sarcastic, respectively	Naïve Bayes	**AMLA**
30.	Parde and Nielsen (2018)	Polarity: positive and negative tweets	Naïve Bayes	**AMLA**
31.	Bouazizi and Ohtsuki (2018)	polarity: positive, negative or neutral	Random forest	**AMLA**

Nota. Tabla resumen de estudios y algoritmos utilizados para identificar la detección del sarcasmo. Fuente Sarsam et al. (2020).

Ahora bien, los autores también plasman la clasificación en el esquema AMLAs representado en la Figura 14.

Figura 14

Clasificación AMLA



Nota. Gráfico de clasificación de AMLAs para la detección del sarcasmo (*trolling*), Fuente Sarsam et al. (2020).

Los resultados del análisis revelan que SVM fue el modelo más utilizado AMLA para la detección del *trolling* en Twitter, adicionalmente se descubre que la combinación de una red neuronal convolucional CNN ofrece también una alta precisión.

Algo ya mencionado en este estudio es el problema que el *trolling* genera en la salud mental de las personas, el *trolling* y la depresión probablemente son dos términos que están

relacionados. Los estudios para detectar el *trolling* son variados y han cobrado mayor impulso; sin embargo, el logar identificar la depresión es algo que generaría mucho valor utilizando las técnicas y modelos que se han revisado en este documento. Ganguli et al. (2020), realizan un estudio de este impacto y forma de evitarlo.

Ganguli et al. (2020) analizan las publicaciones de los usuarios en Twitter para identificar signos de alerta en los que se muestre algún rasgo de depresión o no, para ello se apoyan en un cuestionario compuesto con 20 preguntas y cada pregunta tiene una escala de 4 puntos del 0 al 3 y la sumatoria de estas preguntas sirve como indicador para determinar la presencia o ausencia de la depresión, también fueron extraídas algunas características de los usuarios como palabras afectivas o no y re-tuis entre algunos datos adicionales.

De la misma manera también se analiza la red de Facebook de la cual se obtiene un conjunto de datos de 600 atributos en base a los likes (me gusta) sobre los cuales se hace el análisis, los resultados indican que los algoritmos de SVM, NB, ME (clasificador experimental) aplicados sobre las publicaciones de Facebook entregan un mayor nivel de rendimiento como se expone en la Tabla 24.

Tabla 24

Resultados obtenidos

Paper Reference	Dataset	Algorithms	Performance Measure
[45]	Facebook	SVM, NB and ME	91%
[48]	Twitter	I-DA and SVM	69%
[50]	Twitter	SVM	85%

Nota. Rendimientos obtenidos de los algoritmos SVM, NB, ME, LDA, Fuente Ganguli et al. (2020).

Desde otro punto de vista Lundberg et al. (2020) exponen en su trabajo la injerencia de *trolls* y sus operaciones en sembrar caos en las sociedades occidentales. Su trabajo inicia con un relevamiento los datos al contenido generado por los trols mediante un análisis a 3.5 millones de mensajes obtenidos de Twitter en el 2018. Para filtrar los datos necesarios para el análisis

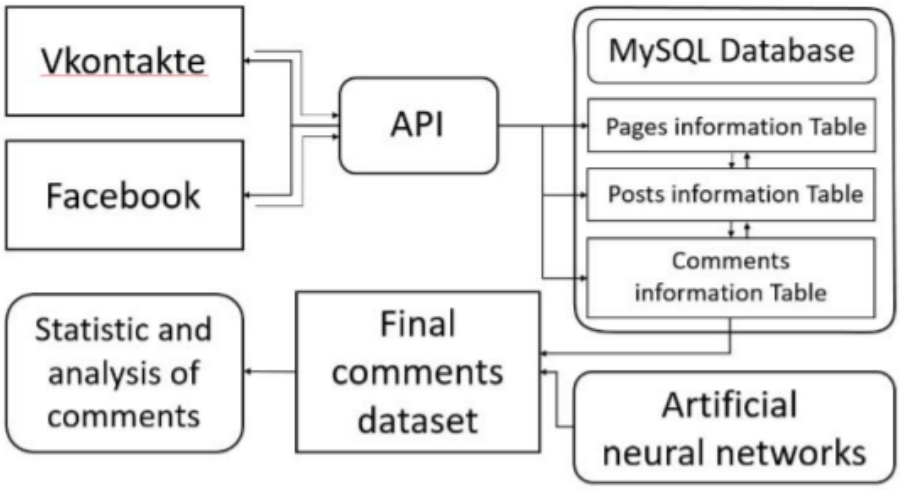
se utiliza una aplicación de aprendizaje automático que permite seleccionar los mensajes exactos, aunque en el estudio se muestran graficas de análisis importantes no se verifica las características de la aplicación y los modelos utilizados, así como su precisión, sin embargo, se menciona la utilización de un algoritmo creado en el 2019 por Lundberg con un enfoque independiente del idioma para detectar tuits generados por un *bot*.

En los experimentos realizados indican que el mejor algoritmo fue J48-HP el cual tuvo un valor de precisión del 0,998 y un Recall del 0,920 con un margen error estuvo en 1,84%.

Ya para el 2020 se han mostrado varios artículos que muestran primero el continuo crecimiento del *trolling*, la injerencia de *trolls* en las decisiones humanas y las técnicas que se utilizan para detectar esta realidad. El estudio que se muestra a continuación ha considerado el mencionar abiertamente el almacenamiento de información usando una base datos “MySQL”, el estudio mencionado es el de Gamova et al. (2020) el cual plantea desde el inicio una arquitectura para abordar esta tarea, como se visualiza en la Figura 15.

Figura 15

Marco de referencia



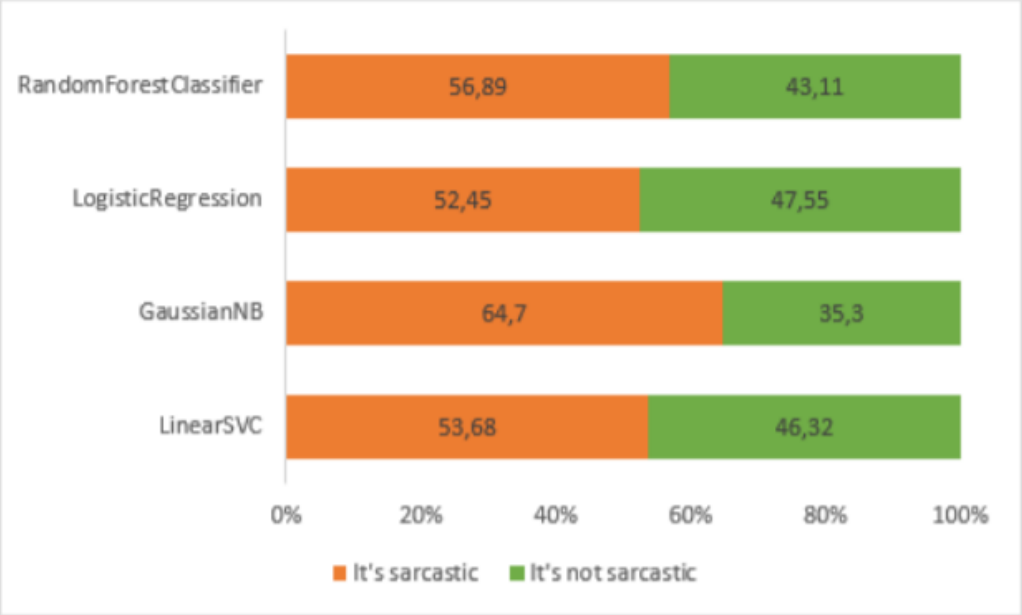
Nota. Marco de referencia para el análisis de datos. Fuente Gamova et al. (2020).

Uno de los retos es conseguir la mayor parte de los datos para el análisis, esto se consigue mediante una API de conexión a redes sociales como Vkontakte, Facebook y Twitter, de donde se extraen los datos.

Posteriormente, utilizando algoritmos como Random Forest, Logistic Regression, Gaussian NB se procesan los datos obtenidos, el resultado muestra una clasificación del sarcasmo que excede la media y el mayor valor se sitúa en 56.89 utilizando el algoritmo de Random Forest como lo muestra la Figura 16.

Figura 16

Resultados obtenidos



Nota. Resultados de los algoritmos LR, GNB, Linear SVC utilizados para la detección del Sarcasmo. Fuente Gamova et al. (2020).

Los autores concluyen que los resultados obtenidos no tienen una precisión significativamente buena, lo cual solo se logra con la ayuda manual, la percepción humana y la redacción semántica son factores muy importantes al momento de clasificar.

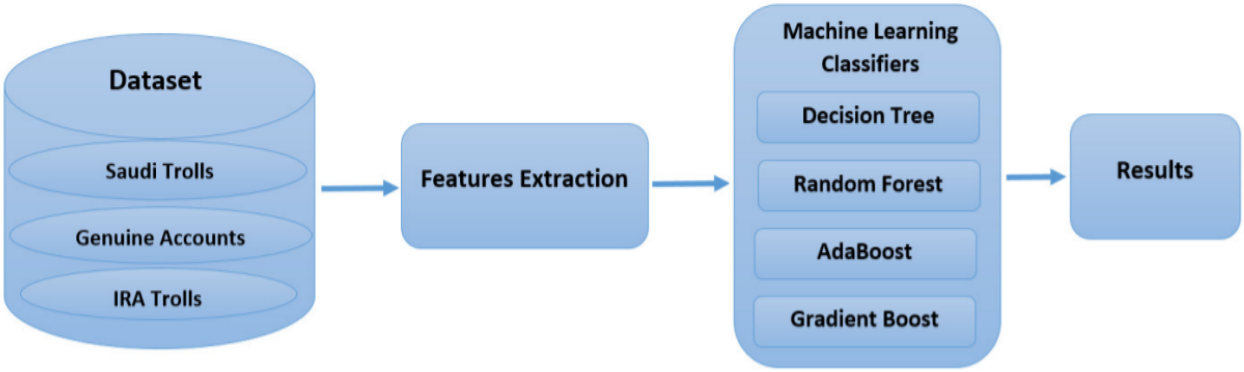
Investigadores como Alhazbi, S. (2020) realizan un estudio sobre la influencia del gobierno en la opinión pública a través de campañas de desinformación en las redes sociales, este es un fenómeno que no solamente ha afectado a personas comunes sino también a grupos políticos, generando desconfianza interna y externa. El autor comenta la necesidad de desarrollar mecanismos que ayuden a identificar las cuentas *trolls* para mitigar esos impactos en los procesos políticos.

El análisis del estudio inicia con la obtención de datos desde las cuentas de Twitter, para que posteriormente sean procesadas y analizadas, para ello propone la utilización de cuatro modelos de clasificación estos modelos se basan en los algoritmos de Decision Tree, Random Forest, Adaboost y Gradient Boost.

Se plantea el siguiente modelo de trabajo que permita comprender el análisis ilustrado en la Figura 17.

Figura 17

Marco de referencia



Nota. Marco de referencia utilizado para el análisis de datos. Fuente Alhazbi S. (2020).

Como se había mencionado los datos fueron extraídos de Twitter, estos contenían 3.841 cuentas de las cuales 2286 eran de la Agencia de Investigación de Internet Rusa (IRA); se seleccionaron únicamente las cuentas que se encontraban en inglés adicionalmente se realiza un muestreo aleatorio y se agrega al dataset 171.291 cuentas de los Estados Unidos.

El modelo fue entrenado con datos de *trolls* Saudies revelados por Twitter en el 2019, la precisión de la clasificación fue del 94.4% como se muestra en la Tabla 25.

Tabla 25

Resultados obtenidos

Model	Accuracy	Precision	Recall	F-Score	AUC
Decision Tree	0.927	0.935	0.912	0.924	0.926

Model	Accuracy	Precision	Recall	F-Score	AUC
Random Forest	0.937	0.950	0.918	0.934	0.937
Adaboost	0.923	0.923	0.906	0.919	0.922
Gradient Boost	0.944	0.956	0.927	0.942	0.944

Nota. Rendimiento de los algoritmos utilizados en la detección de los *Trolls* Sauides. Fuente Alhazbi, S (2020).

Los modelos tuvieron la capacidad de identificar a *Trolls* rusos con una precisión de hasta el 72.6% sin entrenamiento en este conjunto de datos.

Se muestra que los resultados de las características de identificación de un grupo y otro pueden variar, la realidad es distinta en cada grupo, por lo tanto, se debe buscar mecanismos para mejorar las clasificaciones como por ejemplo el obtener información de otros países para los procesos de aprendizaje como se muestra en la Tabla 26.

Tabla 26

Resultados obtenidos

Model	Accuracy
Decision Tree	0.721
Random Forest	0.677
Adaboost	0.726
Gradient Boost	0.641

Nota. Resultados obtenidos de los algoritmos DT, RF, Adaboost, GB utilizados en la detección de los *Trolls* de IRA. Fuente Alhazbi S (2020).

La identificación del contenido agresivo en el estudio realizado por Ramiandrisoa (2020) nuevamente plantea la forma de detectar el "*Trolling*". Este mismo enfoque se presenta en estudios similares analizados.

Los autores plantean en su trabajo la utilización de modelos basados en aprendizaje automático supervisado, dividiendo el análisis en dos fases. La primera fase aborda una combinación de

bosques aleatorios y Regresión Logística basada en vectorización de documentos y la segunda fase aborda el Deep Learning “aprendizaje profundo”.

Se menciona que en el 2018 se realiza una tarea compartida sobre identificación de agresiones como parte de un primer taller desarrollado para identificación del *trolling*, donde 30 equipos presentaron pruebas obteniendo una puntuación F-score de 0.65.

El desarrollo tiene como propósito analizar 11.990 comentarios de Facebook sobre los cuales se realiza un proceso de clasificación para determinar los niveles de agresión agrupado en: abiertamente agresivo OAG, encubierto agresivo CAG y no agresivo NAG, adicionalmente se utilizan 1257 tuis como segundo conjunto de prueba como se muestra en la Tabla 27.

Tabla 27

Distribución de datos

Number of	**Train**	**Validation**	**Test**	
			Facebook	**Twitter**
texts (=posts+comments)	11,999	3,001	916	1,257
Overt aggression	2,708	711	144	361
Covert aggression	4,240	1,057	142	413
No aggression	5,051	1,233	630	483

Nota. Distribución de los datos para entrenamiento y validación. Fuente Alhazbi S (2020).

La métrica utilizada es el F1 ponderada, siendo este un indicador de precisión y exhaustividad que también se utilizó en la tarea compartida para la identificación de la agresión y el *cyberbullying* denominada TRAC; los resultados se muestran en la Tabla 28.

Tabla 28

Resultados Obtenidos

System	**Weighted F1**	
	Facebook	**Twitter**
Random Baseline	0.354	0.348
Saroyehun [2]	0.642	0.592

System	Weighted F1	
	Facebook	**Twitter**
Trac-RF_LR	**0.581**	0.409
Trac-CNN_LSTM	0.559	**0.511**
Trac-RF_on1y	0.573	0.397
Trac-LR_only	0.569	0.452

Nota. Resultados de los algoritmos utilizados para el análisis. Fuente Alhazbi, S (2020).

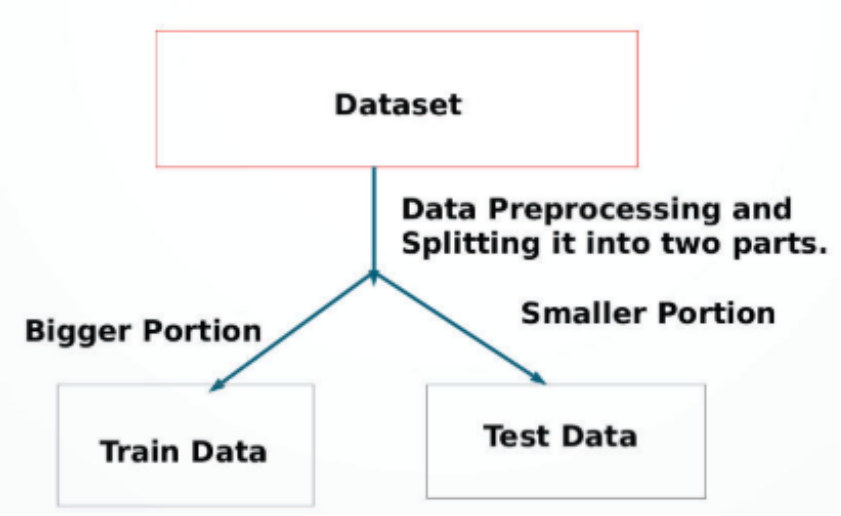
El estudio concluye con un análisis al TRAC comentando sus desafíos y resultados, finalmente se menciona el realizar una nueva investigación sobre los modelos de Deep Learning utilizando diferentes arquitecturas como las redes de atención jerárquicas *Hierarchical Attention Network* para generar modelos de mayor rendimiento.

Los investigadores Singh et al. (2020) miembros del Instituto de Información y Tecnología de la India realizan un planteamiento similar al que muchos estudios analizados, ellos abordan el troleo y el acoso en línea tratando de determinar comentarios de odio en sus diversas formas. La investigación también nombra métodos usados por Google y Jigsaw para la identificación de *trolling* utilizando una aplicación llamada Perspective alojada en los servidores de Google para distinguir actos de provocación y maltrato. Esta *application programming interface* (API) utiliza modelos de AI para puntuar en tiempo real a los comentaristas, y aunque la API fue de mucha ayuda no fue muy utilizada por los usuarios.

En este sentido se propone clasificar comentarios extraídos de la fuente Kaggle Competition, el dataset contiene discusiones extraídas de Wikipedia que han sido etiquetados para analizar comportamientos clasificados en algunas categorías como tóxico, toxico severo, obsceno, amenaza, insulto, odio de identidad; esta información se divide en 80% para entrenamiento y el 20% para verificación de la precisión tal como se muestra en la Figura 18.

Figura 18

Marco de referencia



Nota. Esquema de separación de comentarios. Fuente Singh et al. (2020).

Para el análisis y clasificación de los datos se utiliza Python con algunas bibliotecas como Pandas, Scikit-Learn, y para la ejecución se utiliza modelos como OneVsAll, Logistic Multi-Level, Bi-LSTM y Bi-Gru, los cuales han entregado los resultados que se muestran en la Tabla 29.

Tabla 29

Resultados obtenidos

Model	Accuracy
Logistic OnevsAll	89.99
Logistic Multi-Label	85.44
Attention Mechanism	96.39
Bi-LSTM with Bi-GRU	97.89

Nota. Resultados obtenidos de los algoritmos utilizados para el análisis. Fuente Singh et al. (2020).

Los estudios concluyen, que Bi-LSTM y Bi-GRU tuvieron la mejor precisión al momento de clasificar el cual fue usado más adelante, aunque no se pudo probar con datos en vivo.

Retomando nuevamente la temática sobre la injerencia política, se dice que la información y la desinformación forman parte de las estrategias comunicacionales de los gobiernos y agrupaciones sociales por lo tanto Smitha et al. (2021) muestran en su investigación actores de la desinformación en las redes sociales. Este trabajo propone un análisis a los datos de Twitter recopilados durante las elecciones francesas en el 2017 y también de las cuentas denominadas Operaciones de Influencia en Línea (IO) divulgadas por Twitter entre los años 2007 a febrero del año 2020.

El dataset utilizado para este análisis tiene información de 50.000 cuentas (IO) correspondientes a 17 países, incluidas cuentas de *trolls* y *bots* según señala el estudio, para analizar tal volumen de datos el autor propone un marco de referencia con 5 fases cono se expone en la Figura 19.

Figura 19

Marco de referencia

Nota. Marco de referencia propuesto para el análisis de datos. Fuente Smitha et al. (2021).

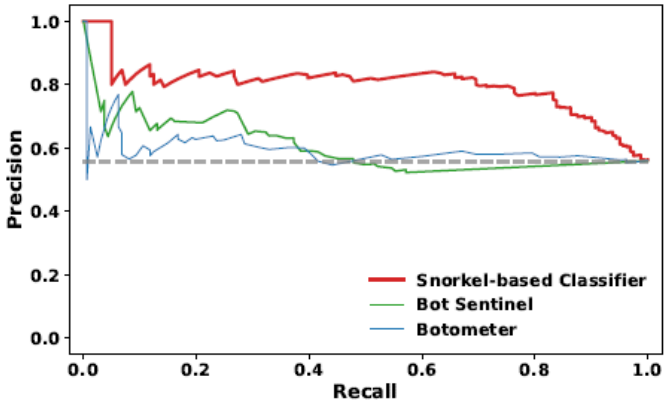
Los datos recopilados para el análisis son extraídos de 28 millones de publicaciones en Twitter y cerca de 1 millón de cuentas que hacen referencia a las elecciones francesas del 2017. Usando el API de Twitter se realiza la colecta de datos basados en palabras clave como #Macron y #LePen, cuentas, y filtros específicos sobre asuntos relacionados a las elecciones y narrativas hostiles que se presume puedan afectar a los candidatos.

Para el desarrollo del modelo de clasificación se utilizó las bibliotecas de código abierto de Scikit-learn y Snorkel que son librerías de Python usadas para realizar análisis predictivo, estas incluyen clasificadores de clusterización etc.

Se presenta a continuación los clasificadores utilizados y sus resultados ilustrados en la Figura **20**.

Figura 20

Resultados obtenidos

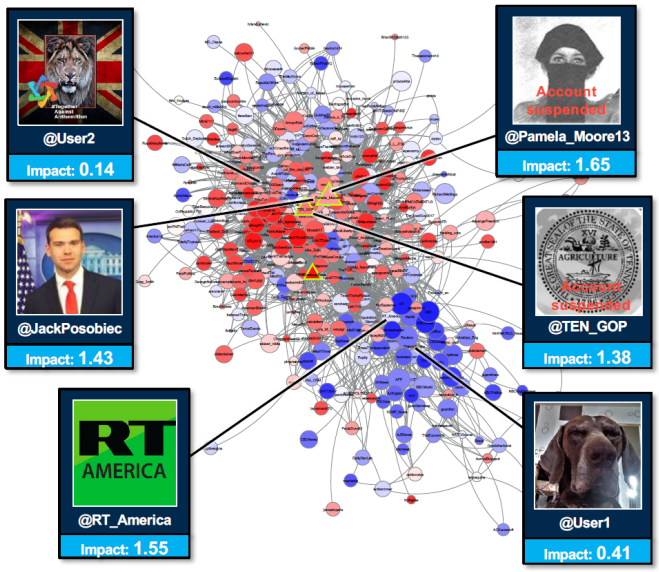


Nota. Resultados de clasificadores. Fuente Smitha et al. (2021).

Los resultados de la clasificación muestran IO que se llevaron a cabo durante las elecciones francesas del 2017 proporcionando alertas a posibles cuentas de operaciones en línea como se expone en la Figura 21.

Figura 21

Resultados obtenidos en forma gráfica



Nota. Red de impacto y cuentas por impacto - Las cuentas IO se resaltan en los triángulos, Fuente Smitha et al. (2021).

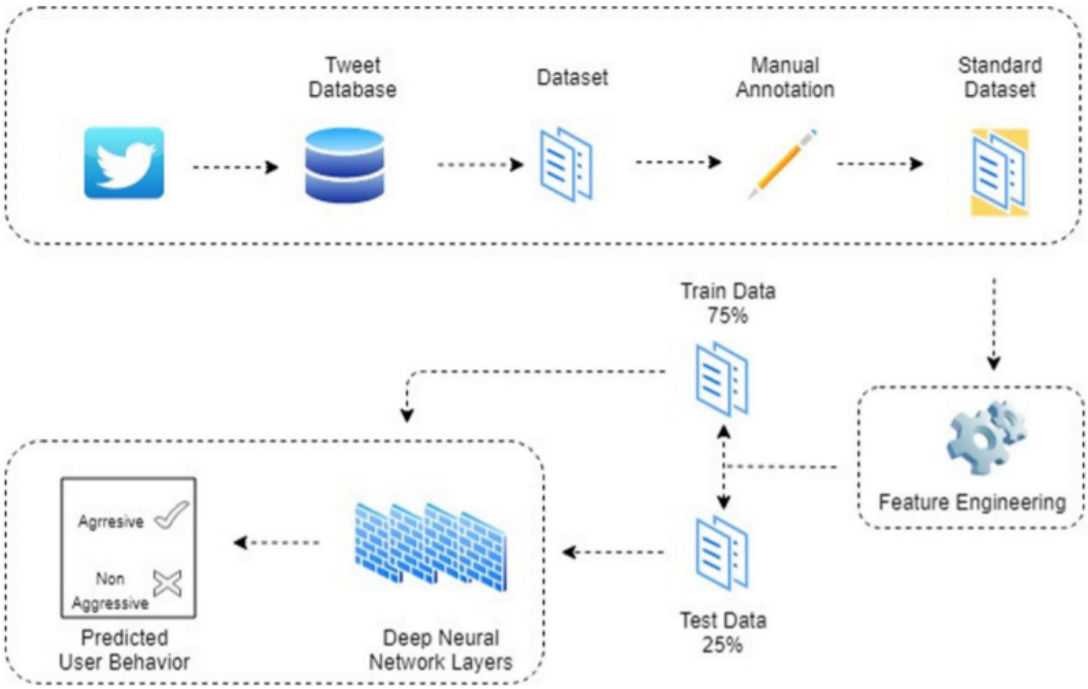
El estudio concluye que el sistema detecta cuentas de IO con una precisión del 96%, un 79% de Recall y un 8% de error, datos bajo la curva con un 96% de precisión superiores a los detectores online Botometer y Bot Sentinel en un 20%.

Un estudio reciente para la detección del ciberacoso es desarrollado por Sadiq et al. (2021). Para estos autores la detección del odio y de la agresión es importante y en su artículo abordan la detección automática de las agresiones sobre la red social Twitter, para el desarrollo de este objetivo utilizan un dataset que contiene cuentas de *trolls*.

El análisis utiliza *multiplayer perceptrón* que contiene características con una combinación de redes neuronales profundas, el análisis comienza desde la construcción de una arquitectura que muestra el inicio, proceso y el resultado como se puede evidenciar en la Figura 22.

Figura 22

Marco de referencia



Nota. Marco de referencia propuesto para realizar el análisis de datos. Fuente Sadiq et al. (2021).

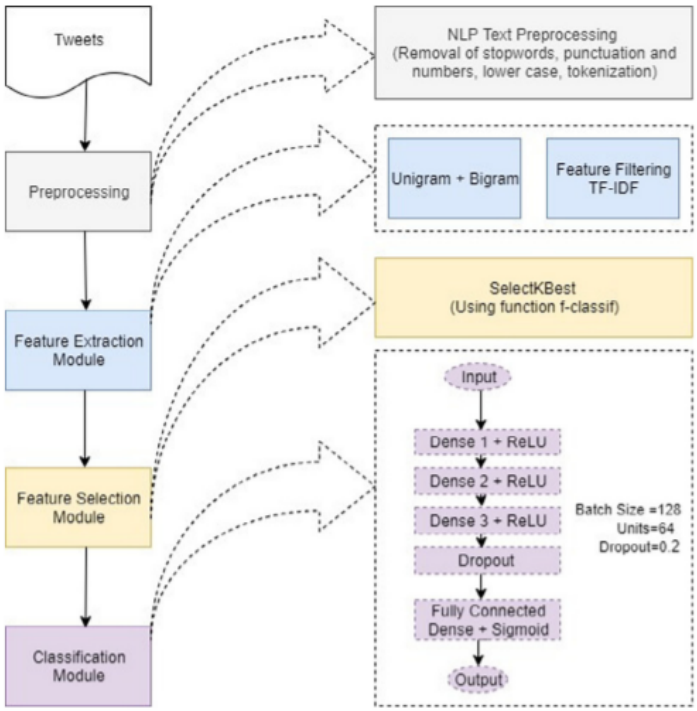
Un hecho particular que el autor nombra en su artículo está relacionado con Tay, un *bot* lanzado por Microsoft via Twitter en el 2016. Tay tiene la capacidad de aprendizaje automático para

identificar el contenido de tipo *trolling,* pero paradojamente este fue atacado u suado por *trolls* para publicar tweets ofensivos y racistas usando un lenguaje fuerte. Microsoft tuvo que detener la ejecución de este *bot* después de 16 horas de su lanzamiento lo que significa que incluso para compañías grandes la detección de *trolls* es una tarea compleja adicionalmente el reto de análisis del lenguaje natural, idioma, modismos y cultura lo complican aún más.

Se detalla en el artículo lo difícil de procesar el lenguaje natural ya que este carece de expresiones faciales o tonos de voz y esto torna complejo cualquier análisis, no obstante, la utilización los algoritmos de aprendizaje profundo conseguirían solucionar estos problemas. Los autores del análisis sugieren utilizar modelos que usan Deep Neural Network DNN dividiendo el desarrollo en tres módulos como: característica de extracción, característica de selección y de clasificación. También como se menciona anteriormente se utiliza una arquitectura de red neuronal simplificada basada en capas completamente conectadas con el multiplayer perceptrón, el esquema propuesto se presenta en la Figura 23.

Figura 23

Modelo



Nota. Representación del modelo DNN. Fuente Sadiq et al. (2021).

El análisis muestra la evaluación de desempeño de cada modelo para la detección de los tuits que tienen agresividad y no agresividad mostrada en la Tabla 30. Se presenta también los resultados de la exactitud "Accuracy", precisión "Precision", recuperación "Recall", y el score F1 de los modelos propuestos.

Tabla 30

Resultados obtenidos

Methods	**CNN-LSTM (Using word embedding)**	**CNN-BiLSTM (Using word embedding)**	**Multi-Layer Perceptron (MLP) (Using word embedding)**	**Proposed approach (Using TF-IDF (unigram+Bigram))**
Accuracy %	86%	87%	87%	92%
Precision%	93%	93%	85%	90%
Recall%	78%	83%	93%	90%
F1 Score%	85%	88%	89%	90%

Nota. Resultados obtenidos de los diferentes algoritmos utilizados. Fuente Sadiq et al. (2021).

Los resultados indican que el modelo funciona y detecta comportamiento agresivo en los tuits con un 92% de precisión.

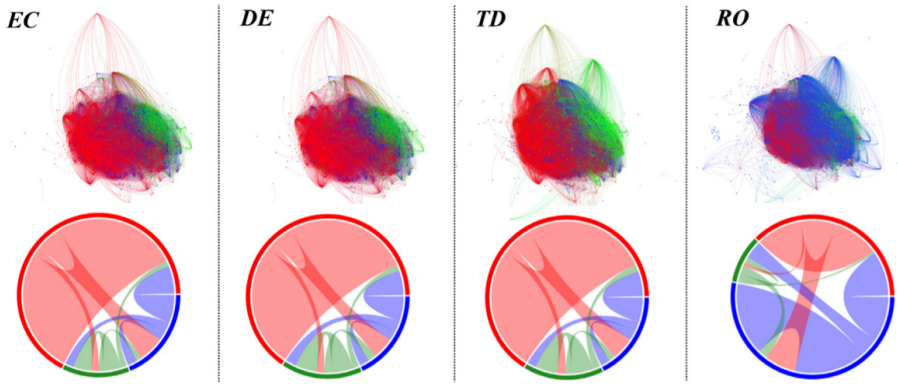
Las discusiones de los usuarios desarrolladas en las redes sociales respecto a los temas políticos en países y relacionados a las elecciones y sus candidatos han sido comunes y quizá repetitivas a lo largo de esta revisión sistémica en este sentido Lai et al. (2019) exponen en su trabajo de investigación el analizar el referéndum del 2016 sobre la reforma de la constitución italiana y los mensajes de *bots* y *trolls* en este ámbito y la forma de identificarlos.

Para ello se obtiene información de Twitter mediante el API Rest de Twitter, obteniendo más 2 millones de tuits completos a los cuales les aplicó un filtro de hashtag, el conjunto de datos fue recuperado mediante la API Twitter Stream API, del 24 de noviembre al 7 de diciembre, generando 900k tuits.

El análisis realizado se centra en las redes de seguidores, como lo muestra la Figura 24, la relación de la red basada en amigos es dividida en cuatro fases tomando en cuenta que la información fue tomada después del referéndum de la reforma a la constitución italiana.

Figura 24

Gráfico de clasificación



Nota. Arriba el diseño atlas, abajo el diagrama de cuerdas, los colores representan, verde a favor, rojo en contra y azul ninguno. Fuente Lai et al. (2019).

Un algoritmo utilizado para este análisis es el SVM el cual ha generado *scores* de clasificación interesantes.

El estudio concluye que, en base a los tuits generados, los usuarios tienden a expresarse de manera más neutral luego del referéndum. Adicionalmente se menciona la importancia del modelo de redes de comunicación y cómo los usuarios expresan sus diferentes posturas.

Más estudios del comportamiento humano se gestan cada año en el mundo, la tecnología aporta hoy en día herramientas para el análisis de diferentes realidades que ayudan a determinar el comportamiento obsceno y causante de caos en las redes sociales. Es así como Borzemski, L. (2019) propone también un método de análisis a la información que desinforma y causa conflicto, el contenido analizado incluye ciberacoso e insultos.

Se plantea el uso del algoritmo de machine learning que en su concepto base tiene capacidades de aprendizaje y un nivel de respuesta automático, para el desarrollo de este trabajo se usan para los siguientes algoritmos:

- **K-nearest Neighbors:** Modelo de clasificación con vecinos cercanos.
- **Naïve Bayes Classifier:** Clasificador que usa el teorema de bayes.
- **Decision Trees:** Modelo de árboles de decisión.
- **Logistic Regression:** Modelo de combinaciones lineales de variables independientes.
- **Support Vector Machine:** Método que separa un espacio muestral en dos o más clases con el margen más amplio posible.

Los datos se obtuvieron de dos fuentes, la primera fuente fue los comentarios realizados en Wikipedia y extraídos mediante figsare.com de la cual se entregaron algunos comentarios etiquetados; y, la segunda fuente es Kaggle que en conjunto sumaron 400k. de estos comentarios 40k fueron tóxicos.

Posterior a la ejecución de los modelos, el autor muestra una tabla con los hallazgos más significativos identificando con negrita los valores que obtuvieron mayor precisión en el análisis descrito en la Tabla 31.

Tabla 31

Resultados Obtenidos

ML	Tf-Idf weighting			Average		
method	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
KNN	46.23	59.62	52.08	41.46	50.80	45.62
NB	**80.72**	57.28	67.01	**73.07**	58.97	65.12
DT	70.35	74.83	72.52	70.44	74.70	72.51
LR	49.98	**84.92**	62.93	54.86	**83.36**	65.14
SVM	73.82	66.66	**70.06**	**72.57**	74.40	**73.34**

Nota. Resultados de los algoritmos utilizados KNN, NB, DT, LR, SVM. Fuente Borzemski, L. (2019).

Este estudio concluye que los modelos LR y SVM son los que entregaron mejor precisión en los análisis, y SVM tuvo el mejor resultado según la medida F1, por tanto, este modelo se utilizó para construir una aplicación.

Ya en los últimos años se han desarrollado estudios que básicamente tratan erradicar el *trolling* mediante el reconocimiento de comentarios ofensivos utilizando modelos de aprendizaje automático. Es por eso que MacHova et al. (2021) exponen esta problemática. Inician realizado encuestas para determinar la identificación del contenido de *trolling* o usuarios *trolls* en redes sociales, llegando a tener algunos datos importantes como: el 74% de los encuestados fueron víctimas o testigos del *trolling*.

Los investigadores comentan las posibilidades de la existencia de *Trollbots*, que son automatismos de menor complejidad comparados con los *bots*, que en algunos casos pueden ser diseñados para influir a las personas y sesgar sus creencias políticas.

Adicionalmente se aborda la existencia de una herramienta llamada *Haters News* que determina el número de mensajes de odio en redes sociales, principalmente analiza las redes de Twitter y Reddit. A esto se suma otro automatismo llamado Cluebot implementado en el 2007 el cual usa métodos de aprendizaje automático para detectar actos de vandalismo en Wikipedia. El modelo que se utiliza actualmente es Naïve Bayes, en teoría Cluebot puede detectar el 90% de cambios perjudiciales en las contribuciones en esta red

Analizando el estudio de MacHova et al. (2021) este se enfoca en examinar los comentarios de la red social Facebook, mencionan que el extraer datos de esta red social se complica por el último acontecimiento de uso indebido de datos por parte de **Cambridge Analítica**, por tal motivo se decide utilizar un servicio en Internet que permite exportar comentarios de redes sociales. Los datos fueron recopilados durante un mes desde el brote de la pandemia en el país de Eslovaquia y en total se obtuvieron 2500 comentarios

Para el análisis se utilizó tres modelos SVM, Multinomial Naïve Bayes MNB y LR, en los primeros resultados se identifica que SVM tuvo el mejor rendimiento su precisión fue alta mientras que el Recall fue bajo, tal como se muestra en la Tabla 32, pero el modelo (MNB) tuvo un resultado más estable en el Recall.

Tabla 32

Resultados obtenidos

Bag of Words		**Measure of Effectiveness**		
		Precision	**Recall**	**F1-rate**
SVM	Troll	1,00	0,08	0,15
	Non-troll	0,56	1,00	0,72
MNB	Troll	0,63	0,92	0,75
	Non-troll	0,89	0,53	0,67
LR	Troll	0,67	0,89	0,76
	Non-troll	0,90	0,69	0,78

Nota. Resultados de los algoritmos aplicados. Fuente MacHova et al. (2021).

Como conclusión se indica que no se descubrieron muchos *trolls* en las discusiones, aunque el modelo más adecuado fue el MNB llegando a obtener un porcentaje de efectividad en el Recall del 92%, el enfoque a futuro sería utilizar redes neuronales profundas.

3.2 Aplicaciones desarrolladas para detectar el *trolling*

Hasta este momento se han analizado los modelos supervisados que puedan identificar rápidamente el *trolling*, no obstante, existen un gran número de trabajos que se enfocan en poner en producción este tipo de modelos dentro de aplicaciones que automáticamente pueden reconocer patrones de *trolling,* estos estudios son analizados y presentados en este apartado tratando de identificar la forma en que lo hicieron y los modelos utilizados.

Los esfuerzos para mitigar el *trolling* son cada vez mayores y precisos, sin embargo, los *trolls* también avanzan en la creación de formas para evadir estos controles, por lo tanto, las empresas han desarrollado mecanismos de bloqueo que identifican palabras abusivas en foros, redes sociales o sitios web. En este aspecto y para promover y mostrar los avances de estos sistemas Bhatt et al. (2018) presentan en su análisis una perspectiva de los resultados y problemas que las tecnologías actuales han enfrentado.

Los autores comentan que una empresa inglesa ha creado la herramienta llamada SMC4 la cual captura los datos entrantes los filtra y bloquea contenido fuerte como insultos, sexismo, racismo, entre otros. Además, aseguran ser la primera empresa en el mundo en desarrollar una aplicación *anti-trolling* la cual se utiliza para bloquear los contenidos en Twitter. Por otro lado, Google en alianza con Jigsaw creó el sistema *anti-trolling* llamado Perspective que es una API que usa el aprendizaje automático para detectar abusos y toxicidad de contenidos en línea, y Microsoft mediante sus sistemas de inteligencia artificial detectan contenido ofensivo para posteriormente bloquearlo.

El estudio concluye que los sistemas muestran protección ante comentarios de tipo *trolling* sin embargo no llega a un cien por ciento de efectividad debido a la complejidad humana en su escritura semántica o el llamado lenguaje natural.

Por otro lado un estudio desarrollado por Li et al. (2018) plantea la problemática del sabotaje refiriéndose a sistemas de apoyo y profesionales que brindan ayuda a las personas que experimentan algún tipo de trastorno como la ansiedad o la depresión y los *trolls* que participan de forma maliciosa y atacan a los voluntarios que aportan y apoyan a las personas con este problema.

El estudio indica que para mitigar esta realidad el sistema denominado TeaFilter detecta el comportamiento no adecuado de algunos usuarios, logrando así tener una precisión del 77.8% en la detección con una tasa de error del 1%.

Desde la un enfoque psicológico los autores Li et al., (2018) exponen al 7 Cup como un servicio web de ayuda psicológica para dar apoyo a personas con trastornos de ansiedad o depresión. Comentan que en el 2013 se convirtió en una plataforma de ayuda importante la misma que tiene hasta ese momento unos 600.000 usuarios registrados generando unas 90 mil conversaciones por semana estos perfiles de usuarios se clasifican en dos tipos los miembros y los usuarios escuchas. El objetivo de este trabajo es identificar y bloquear comentarios violentos o que generen caos.

Para la realización del análisis se toma como punto de partida el obtener del administrador de la plataforma 7 Cup o un usuario escucha los contenidos inapropiados que son bloqueados,

cabe decir que estos comentarios están en lenguaje natural y su sintaxis varía, la Tabla 33 muestra un ejemplo de esos cometarios.

Tabla 33

Listado de comentarios

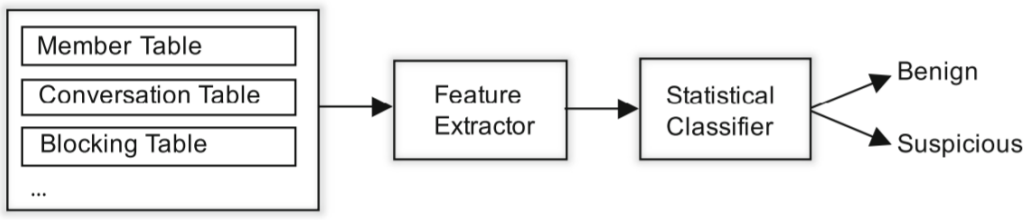
Sexual behavior toward me.
Just here to flirt.
Saying sexual things.
Seemed racially insensitive.
Threatening me.
Very vulgar and abusive language.
Asking personal questions about listeners.
Member requesting too much personal information.
Asking for age, address etc.

Nota. Ejemplo de los comentarios realizados en actividades sospechosas. Fuente Li et al. (2018).

Es sistema TeaFilter se compone de dos fases, la de entrenamiento y la de detección. La fase de entrenamiento aprende de los miembros benignos y los sospechosos mientras que la fase de detección se alimenta de los miembros nuevos que son representados por un vector de características, posteriormente los datos son analizados por el clasificador, en la Figura 25 se muestra la arquitectura del sistema.

Figura 25

Marco de referencia



Nota. Arquitectura funcional planteada del software. Fuente Li et al. (2018).

Respecto a la evaluación el sistema usa tres clasificadores, RF, SVM, y GB Tree, los valores del área bajo la curva AUC se resumen en la Tabla 34.

Tabla 34

Resultados obtenidos

Classifier	**AUC**
Random forest	0.9851
SVM	0.9673
Gradient-boosted tree	0.9766

Nota. Resultados de AUCs para tres clasificadores. Fuente Li et al. (2018).

El estudio concluye y comenta en la importancia del sistema para detectar automáticamente miembros sospechosos en un sitio de apoyo emocional en línea y la aplicabilidad de estos métodos en las redes sociales convencionales.

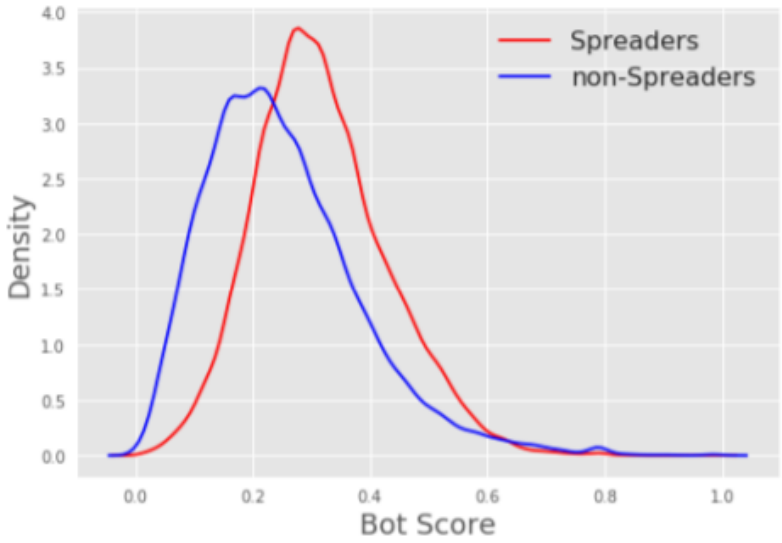
Continuando el análisis del estudio de Badawy et al. (2019), el cual fue señalado en la sección 3.1, donde se hace referencia al Botometer, este sistema analiza un conjunto de más de mil funciones y abarca seis subclases:

- Usuario: funciones de metadatos, incluye número de amigos, seguidores y numero de tuits.
- Amigos: aquí se considera el retuitear, mencionar, ser retuiteado y ser mencionado.
- Red: Se construyen las reglas como, mención y hashtag y redes.
- Temporal: características de las actividades del usuario.
- Contenido: Estadísticas sobre la longitud y la entropía del texto.
- Sentimiento: características como la excitación y la dominancia.

El Botometer se utiliza para etiquetar a los usuarios esparcidores es decir a los que masifican información de los cuales se obtiene un puntaje *bot* para determinar cuáles usuarios pueden ser *bots,* esta información se expone en la Figura 26.

Figura 26

Resultados obtenidos



Nota. Gráfico que muestra el *Score* de clasificación de *bots.* Fuente Badawy et al. (2019).

Las conclusiones del estudio muestran un análisis profundo de los esparcidores de contenido, sus técnicas y el reto de poder detectar estos comportamientos para tener menor sesgo en las discusiones de foros relacionados a la política y así poder tener sociedades más justas.

Vale la pena indicar que la mayoría de los usuarios en Internet utiliza un navegador para poder acceder al contenido de los diferentes sitios, páginas o servicios que ofrecen diversidad de información. El navegador web se ha vuelto esencial y es uno de los medios más usados por las personas para conectarse a la red (Modha et al., 2020).

En este contexto y ante el crecimiento de publicaciones de odio y generación de violencia en las redes sociales Modha et al. (2020) proponen tratar esta problemática con un interesante complemento (*plugin*) para los navegadores, este aplicativo ayudaría al control del *trolling.* En el estudio se comenta la utilización de varios clasificadores como SVM, Logistic Regression, modelo de aprendizaje profundo basado en la red neuronal convulcional CNN, modelo basado en la atención y el modelo de lenguaje pre entrenado BERT propuesto hace poco por Google AI para la clasificación de agresiones.

Los datos se clasifican en etiquetas:

- Abiertamente agresivas (OAG).
- Encubiertamente agresivas (CAG).
- y No agresivas (NAG).

El complemento fue probado en el navegador de Google Chrome el cual genera una etiqueta al inicio de la publicación o *post,* indicado un *score* de agresión en la discusión aportando así con una idea visual como se muestra en la Figura 27.

Figura 27

Gráfico de muestra

Nota. Gráfico que muestra el resultado de la aplicación del Plugin en el navegador de Facebook. Fuente Modha et al. (2020).

La parte más valiosa del estudio se centra en analizar los comentarios de las redes sociales de Facebook y Twitter relacionadas a los posts de los líderes políticos. Para este análisis se obtuvo los dataset denominados TRAC *Trolling* Agression and Ciberacoso los cuales albergan 15k de publicaciones y comentarios de Facebook que posteriormente al ser analizados y procesados se tuvo los siguientes valores indicados en la Tabla 35.

Tabla 35

Resultados obtenidos

TRAC English Dataset

Model	**Facebook engine**			**Twitter engine**		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN with fast Text	0.6996	0.6091	**0.6407**	0.5664	0.5911	0.5520
BiLSTM with attention	0.7059	0.5098	0.5276	0.5766	0.5839	**0.5782**
BERT	0.7156	0.5840	0.6184	0.5623	0.5807	0.5683
Logistic Regression	0.6811	0.5710	0.6046	0.5232	0.5179	0.4890
Support Vector Machine	0.6795	0.5524	0.5902	0.4899	0.4924	0.4853

TRAC Hindi Dataset

Model	**Facebook Hindi**			**Twitter Hindi**		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN with fast Text	0.6261	0.6124	0.6081	0.5206	0.4975	**0.4992**
BiLSTM with attention	0.5904	0.5690	0.5651	0.4861	0.4623	0.4639
BERT	0.6474	0.6216	**0.6182**	0.5101	0.3852	0.3353
Logistic Regression	0.6607	0.6185	**0.6133**	0.3779	0.3731	0.3723
Support Vector Machine	0.5998	0.5886	0.5861	0.3942	0.3874	0.3886

Nota. Resultados de la aplicación de algoritmos. Fuente Modha et al. (2020).

La puntuación F1 ponderada es de 0,64 y 0,62 que se logra en los conjuntos de datos en inglés e hindi del TRAC en Facebook, mientras que en los conjuntos de datos en inglés e hindi de Twitter la puntuación F1 ponderada es de 0,58 y 0,50 respectivamente.

El estudio concluye en que este tipo de problemas se están convirtiendo en el centro de atención de las investigaciones en PNL usando modelos de aprendizaje profundo (Deep Learning), no obstante, el rendimiento sigue estando distante de ser perfecto debido a la complejidad del lenguaje, la violación a las normas lingüísticas, la sintaxis, modismos, coloquialismos que son parte de las culturas en diferentes sociedades.

Jachim et al. (2021) explican en su estudio la creación de un software llamado TrollHunter2020 usado para detectar el *trolling* en tiempo real en la red social Twitter. Este sistema fue utilizado durante y después de las elecciones presidenciales estadounidenses en el 2020. El enfoque de este software es detectar las narrativas de troleo para preservar la línea de respeto en discursos o análisis de determinados temas políticos en la red social.

Muchos datos obtenidos en otros análisis similares se basaban en una temporalidad es decir desde una fecha inicial a una final, sin embargo, ante la realidad electoral que es cambiante a cada momento (Jachim et al., 2021).

TrollHunter2000 obtiene datos en línea de lo que está sucediendo ese momento, datos de hashtags y temas de tendencia como debates de los candidatos, datos de la noche de las elecciones y posterior a estas, toda la información procede de Twitter.

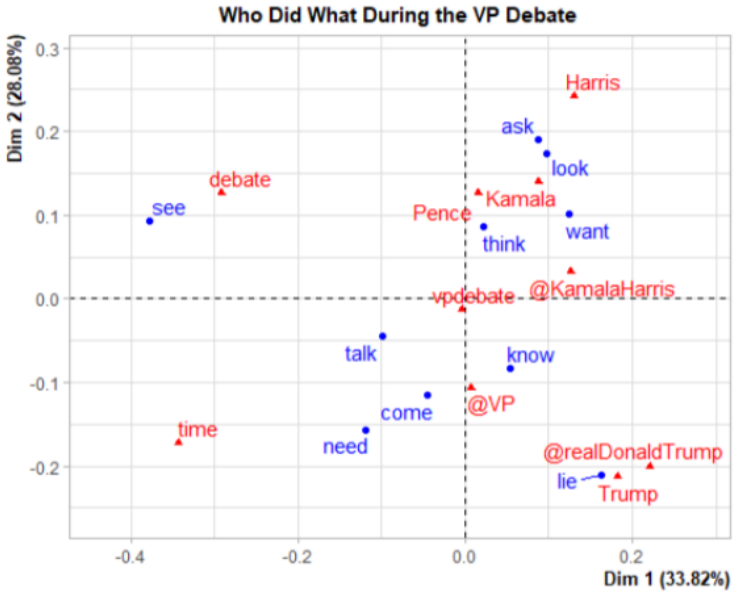
Si bien es cierto en otras menciones de este documento se nombra a la agencia IRA de control de Internet ruso, en este estudio también se lo menciona con un detalle mayor, relacionando a este organismo como un generador de *trolling* político presente en las elecciones de los Estados Unidos en el 2016.

TrollhHunter2000 utiliza la biblioteca de Python Tweepy para acumular los tuis correspondientes a los términos de búsqueda planteados, realiza búsquedas cada 3 minutos a través de las listas ya definidas manualmente, posteriormente usa la biblioteca spacyr para analizar las partes del discurso de cada palabra individualmente de cada tuit y elimina las palabras que no proporcionan mucha información; finalmente crea una tabla en donde deposita

la información de las repeticiones de verbos y sustantivos (variables) a ser analizadas como se muestra en la Figura 28 y Figura 29.

Figura 28

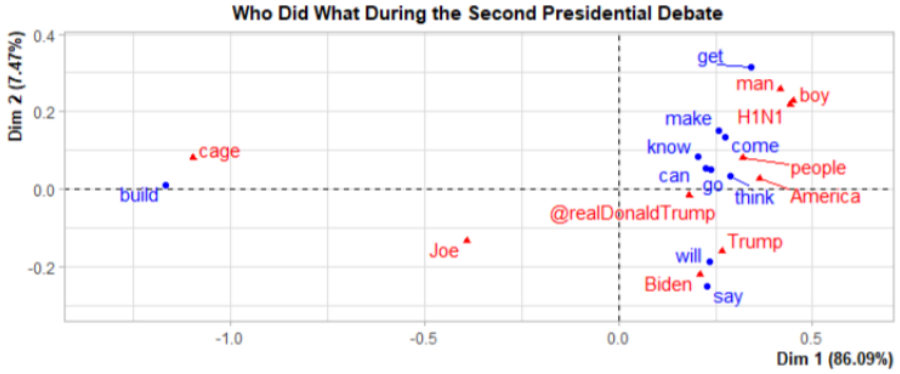
Gráfico de análisis



Nota. Análisis de los verbos y nombres usados en Twitter para describir el debate vicepresidencial en los Estados Unidos. Fuente Jachim et al. (2021).

Figura 29

Gráfico de análisis



Nota. Análisis TOP de los verbos y nombres usados en Twitter para describir el segundo debate presidencial en los Estados Unidos. Fuente Jachim et al. (2021).

Algo importante que el autor presume es la utilización de TrolHunter2000 para propósitos que los determina nefastos, ya que se podría encaminar este análisis a una nueva realidad en el *trolling*, también podría haber una migración de *trolls* a plataformas menos reguladas, aunque quizá con menor llegada a usuarios.

3.3 Análisis del *trolling* mediante el reconocimiento de imágenes

En esta sección se analizan los trabajos que han sido enfocados al reconociendo de imágenes que generan o han generado trolling, por alto contenido viral.

La mayoría de estudios se han centrado en el análisis del *trolling* mediante el reconocimiento del texto utilizando fuentes como Twitter, Facebook u otros, Hitkul et al. (2019) exponen en su artículo el reconocimiento de imágenes propensas al *trolling.* En este caso los autores quieren ser proactivos y no reactivos lo que significa que existirían alarmas antes de publicar una imagen que pudiera causar *trolling*.

Este estudio utiliza un enfoque de clasificación de imágenes utilizando el modelo de red neuronal convolucional CNN y se lo ha divido en tres fases:

1. La primera fase se realiza una clasificación supervisada para las imágenes extraídas.
2. La segunda y tercera fase se evalúa la correlación entre las imágenes (*trolling* o normal).

Para la primera fase mencionada en el numeral 1, se utilizaron 5 algoritmos, los resultados se muestran en la Tabla 36.

Tabla 36

Resultados obtenidos

Model	Validation Accuracy	Test Accuracy
VGG16	63.54	**61.81**
SVM (VGG16 features)	57.79	**61.81**
SVM (VGG19 features)	60.55	58.18
Inception Resnet V2	63.54	55.45

Model	Validation Accuracy	Test Accuracy
Inception V3	**65.62**	53.63
VGG19	63.5	51.81
Custom	54.16	51.81
ResNet 50	52.08	50.90
Zero Rule baseline	50.72	

Nota. Resultados de clasificación de imágenes supervisadas. Fuente Hitkul et al. (2019).

Para la fase 2 y 3 fase mencionadas en el numeral 2, se genera una hipótesis sobre la clasificación, como comparativas de imágenes como se muestra en la Figura 30.

Figura 30

Comparativo de imágenes

Nota. Comparativo de las calidades de las imágenes susceptibles al *trolling*. Fuente Hitkul et al. (2019).

3.4 Detección del *trolling* mediante tensores

Esta sección se menciona como un apartado diferente a las secciones anteriormente citadas en donde se describe una técnica de identificación del *trolling* distinta a las que utilizan algoritmos.

Desde aplicaciones ya desarrolladas, modelos de clasificación, regresión, y uso de tensores, son algunos mecanismos usados en la detección del *trolling*. Los tensores son objetos de

origen matemático que tienen como objetivo guardar valores numéricos con los cuales se puede realizar análisis posteriores (J. Lee et al., 2018).

Por lo tanto J. Lee et al. (2018) en su análisis se proponen un método de detección del *trolling* usando descomposición tensorial compacta (CTD) como propuesta, este método se basa en muestreo interpretable donde el algoritmo encuentra subconjuntos en valores pequeños realizando recorridos a través de un nodo semilla.

Los resultados de este trabajo muestran que si inyectamos un mayor número de datos tipo *trolls* la precisión es mayor, como se muestra en la Tabla 37.

Tabla 37

Resultados Obtenidos

n	Recall	Precision	F1 score
1	1.000	0.200	0.333
3	1.000	0.500	0.664
5	1.000	0.556	0.714
10	1.000	0.833	0.909

Nota. Detección de *trolls* insertados. Fuente J. Lee et al. (2018).

3.5 Resumen de los estudios y la efectividad de los modelos y algoritmos usados

Para una correcta distinción y compresión de los estudios se los clasifica por su año de creación y por la fuente de datos que estos utilizan (redes sociales y otras fuentes).

Estudios que usan datos de fuentes de tipo Redes Sociales.

Los autores **Galán-García et al. (2014)** y **Herrero et al. (2014)** usan los datos de la fuente Twitter, para la clasificación de comentarios y se usa el modelo SVM, el estudio concluye y determina que Polly Kernel PK es un algoritmo que generó mejores resultados.

Dlala et al. (2014) utilizan el algoritmo de k-means aplicado a grupos de usuarios para determinar los trols, no se muestra un valor de efectividad en la clasificación sin embargo se indica que existe éxito en la clasificación.

Tsantarliotis et al. (2016) usan un modelo de árboles de decisión para la clasificación de *trolling*, no se muestra valores precisión que permitan identificar la eficiencia del algoritmo.

Mutlu et al. (2016) analizan los comentarios de *trolling* utilizando como fuente los comentarios de Twitter, para el análisis utilizan algoritmos, KNN, C4.5 y NB, se determina que C4.5 ofrece mejores niveles de clasificación.

Chang et al. (2016) se enfocan en la detección de rumores en Twitter, utilizan un modelo de clusterización, el estudio muestra una parte del código basado en reglas condicionales; no se especifica un resultado final del algoritmo, sin embargo, la regla que mejor precisión muestra es la R3 y R4.

Authors (2017) muestra un estudio basado en la detección de la agresión mediante algoritmos de clasificación como SVM y LR, los autores no revelan a ciencia cierta cual es la fuente de datos utilizada, sin embargo, aseguran haber tenido éxito en la detección de trols utilizando modelos de LR y SVM.

Fornacciari et al. (2018) su estudio se basa en clasificar el *trolling* utilizando la fuente de Twitter, los modelos utilizados son SMO, NB, RF, siendo SMO el que mejor puntuación de exactitud entrega.

Bellan et al. (2018) enfocan su trabajo en la detección del *trolling* mediante la clasificación de comentarios inapropiados en fuentes de noticias, en este caso se utiliza como fuente de datos a Twitter. La información es procesada y analizada utilizando los algoritmos DT, NB, LR posteriormente se determina que el algoritmo que mejor clasificación entrega es el LR con un F1 de 0,64 para comentarios en fuentes profesionales y un F1 de 0,50 para Twitter.

Rasel et al. (2018) utilizan un modelo de clasificación con los algoritmos DT, RF, SVM y LR, el algoritmo que genera una mejor clasificación en el estudio es RF con un 95% de precisión vs LR con un 94%.

Bastos et al. (2018) enfocan su análisis en la detección del *trolling* y sus consecuencias desde un punto de vista político conceptual. No se identifica el uso de modelos y algoritmos, pero si se nombra las fuentes de datos como Twitter. También se comenta el reconocimiento de *bots* como componentes autómatas; toda esta investigación es relacionada al tema del Brexit.

Bhatt et al. (2018) se centran en la identificación del *trolling* mediante una aplicación llamada SMC4 "anti-*trolling*", la lógica de la herramienta no es expuesta y tampoco sus modelos**.**

Li et al. (2018) exponen la detección del *trolling* desde la óptica del sabotaje, la herramienta utilizada es TeaFilter que en su composición utiliza los algoritmos de RF, SVM y Gradient Boosted Tree, siendo el RF el más preciso con un valor de AUC de 0.9851.

Tian et al. (2019) enfocan su estudio en el problema que causa desde el punto de vista psicológico y emocional a las agresiones de tipo trolling. En el análisis no se especifican modelos ni algoritmos para detección del *trolling,* pero se destaca la discusión de la problemática y sus posibles consecuencias.

Kellner et al. (2019) comentan que su análisis se realiza en base la información extraída de Twitter, en el desarrollo se utilizan 7 algoritmos, GB, RF, AdaBoost, SVC, LR, Linear SVC y Knieghbors; el resultado indica un mayor rendimiento en el algoritmo de GB, el valor es F1 0.89 con un 3% de margen de error y AUC 0,976 con un 1% de margen de error.

Luqyana et al. (2019) mencionan la necesidad de erradicar el *trolling* en la red social de Twitter de la cual se extraen los datos para el análisis, aunque no menciona la forma de extracción si se identifica el volumen de los datos. Se comenta que en estudios anteriores existe un problema denominado desbalance de datos que es corregido en este estudio con KNN-Undersampling para luego utilizar el algoritmo de clasificación SVM; el estudio concluye en que existe una ligera mejora al usar un balanceador de datos.

Marouf et al. (2019) centran su trabajo en la identificación de *trolls* que realizan comentarios de desprestigio a personalidades, utilizan para su análisis datos de comentarios y los estados de "Me Gusta" "Likes", de la red social Facebook; para el desarrollo utilizan algoritmos como SVM, MNB, J48, llegando a obtener el mejor puntaje de precisión al utilizar MNB con un porcentaje de 98.6%.

Badawy et al. (2019) usan datos de la red social de Twitter y realizan su análisis utilizando el algoritmo de Gradient Boosting que es un clasificador que tuvo el mejor rendimiento, obteniendo valores entre el 85% y 96% de AUC.

Lai et al. (2019) analizan la información de redes sociales generada en el referéndum italiano, el identificar tendencias políticas es el foco central, para ello se utiliza el algoritmo de SVM, el cual entrega diferentes lecturas respecto a su precisión, no se determinan valores de resultado.

Hitkul et al. (2019) exponen en su estudio que la detección de imágenes sensibles podría desencadenar en acoso y el objetivo de los autores es mitigar esta realidad, para lo cual utilizan 5 algoritmos, VGG16, VGG19 en conjunto con SVM, Inception V2 y V3, RestNet50; en las pruebas de precisión se obtiene valores de 61.81% para VGG16 y SMV-VGG16.

Im et al. (2020) abordan la detección de *trolling* en la política, utilizan para este análisis algoritmos de RL, DT y ADT; en su análisis determinan que el mejor algoritmo para clasificación fue el ADT con una precisión del 94% Recall de 88%.

Dewhurst et al. (2020) en su estudio utilizan los datos obtenidos en la red social de Twitter, para el análisis se utiliza los modelos como No-U-Turn Sampler similares al Gaussian Process Bayesian.

Islam et al. (2020) exponen que su trabajo lo realizan recopilando los datos obtenidos de las redes como Facebook, Twitter y TikTok, se analizan y procesan los datos para la clasificación del *trolling;* y, los algoritmos utilizados son el SVM, RF, DT y NB; siendo el SVM el que mejor clasificación muestra con una precisión del 81%.

Sarsam et al. (2020) este estudio es muy importante ya que agrupa los diferentes algoritmos utilizados en otros trabajos para la detección del sarcasmo. Para ello se genera un puntaje entre dos grupos: los AMLA y los CMLA, los algoritmos del grupo AMLA obtienen mejores rendimientos entre ellos SVM, LR, NB y RF, aunque no se determina el atributo del puntaje, si se determinan los valores respecto a los grupos.

Lundberg et al. (2020) analizan la injerencia de los *trols* en la red social de Twitter, para detectar este comportamiento se utilizan algunos algoritmos de clasificación; y, se determina que el mejor algoritmo fue el J48-HP, reporta una precisión del 99% y un Recall del 92%.

Gamova et al. (2020) centran su estudio en la detección del *trolling* como objetivo, utilizan para este fin varios algoritmos como RF, LR, Gaussian NB y Linear SVC; sin embargo, los

resultados no muestran una alta precisión y tampoco se puntualiza un valor de puntuación definido para cualquiera de los algoritmos.

Alhazbi S. (2020) comentan que medir la influencia de los gobiernos en las redes sociales como Twitter es el enfoque de este trabajo, para lo cual se utilizan los algoritmos como DT, RF, Adaboost y Gradient Boost, la precisión de los algoritmos varía de acuerdo con cada dataset. Los valores registrados para Adaboost es 72% y Gradient Boost es de 94%.

Ramiandrisoa (2020) abordan la detección del *trolling*, para ello el equipo de estudio utiliza un modelo supervisado con algoritmos combinados como RF y LR basada en vectorización de documentos; la segunda fase aborda el deep learning "aprendizaje profundo". Los resultados indican un mejor score en RF + LR con un valor de F1 58% para el set de datos de Facebook y para Twitter con el algoritmo CNN + LSTM un valor F1 de 51%.

Modha et al. (2020) comentan que ya en los últimos años, las técnicas no han variado mucho, el trabajo de del equipo de Modha utiliza algoritmos como SVM, LR y modelo de aprendizaje profundo basado en la red neuronal de convolución CNN. Su análisis se desarrolla en modo real y su enfoque es la detección de *trolls* utilizando comentarios de las plataformas Twitter y Facebook. El resultado indica que se obtuvo los mejores resultados para el modelo de CNN con valor score F1 del 64%, otros modelos le siguen como el de LR.

Singh et al. (2020) exponen que al igual que otros investigadores han centrado sus esfuerzos para erradicar los comentarios negativos de los *trolls*. Este trabajo propone la identificación de comentarios nocivos utilizando minería de datos, usando para ello las fuentes de datos de Kaggle y Wikipedia y para el análisis se usa la herramienta Python y modelos como OneVsAll, Logistic Multi-Level, Bi-LSTM y Bi-Gru. Sus conclusiones determinan que Bi-LSTM y Bi-Gru son los algoritmos que mejor rendimiento han entregado en la clasificación de comentarios nocivos.

Smitha et al. (2021) muestran un análisis de identificación del ciberacoso en el contexto político, para ello utilizan dos algoritmos para la correcta clasificación de los actores de desinformación o "*trolls*" como RF, Extra Trees y xgBoost; aunque el estudio muestra la selección de otros algoritmos, al final el estudio destaca los mencionados como algoritmos con mayor potencial tomando en cuenta que son combinados (RF/ET)

Sadiq et al. (2021) en su estudio analizan la identificación del odio y el ciberacoso en la red social Twitter, los autores proponen una clasificación de los agentes que generan *trolling* mediante la aplicación de redes neuronales profundas utilizando los modelos de CNN-LSTM, CNN-BiLSTM, Multi Layer Perceptron (MLP) y (TF-IDF) (unigram+Bigram). Los *scores* de los algoritmos son variados no obstante si hablamos de precisión y F1 el mejor puntuado es modelo de CNN-BiLSTM con valores de 93% y 88% significativamente

Jachim et al. (2021) presentan en su estudio la utilización de un software para la detección de usuarios *Troll* en publicaciones de redes sociales relacionadas a la política; El software se llama TrollHunter y utiliza datos de la red social Twitter para sus análisis; la documentación el software no detalla los algoritmos utilizados, pero si varios niveles de puntaje sobre las clasificaciones de *troll* y no *troll*.

MacHova et al. (2021) plantean la detección del *trolling* y para lograr esto utilizan datos de Facebook para posteriormente aplicar varios algoritmos como SVM, NB y LR, siendo el NB el que mejor resultado arroja. Se menciona que todavía no se logra obtener un porcentaje significativo en la detección del *trolling* y se sugiere seguir probando modelos como redes neuronales profundas.

Estudios que usan datos de fuentes externas, diferentes a las de redes sociales.

C. M. Lee et al. (2001) no especifican en su estudio un resultado asociado a un tipo de algoritmo o modelo utilizado, el enfoque es la determinación de los algoritmos que más han generado errores en el modelo de clasificación de emociones, lo que si se resalta es que hace 20 años ya se iniciaba a tratar los temas relacionados a la detección del ciberacoso.

Ververidis et al. (2004) analizan que el modelo más utilizado en su trabajo es el bayesiano, el cual fue analizado bajo una matriz de confusión con una correcta clasificación del 56%. Se comenta que es necesario buscar otros mecanismos de clasificación que brinden mayor efectividad o quizá mejorar los conjuntos de datos analizados.

De-La-Pena-Sordo et al. (2015) estos investigadores clasifican los comentarios de *trolls* mediante la utilización de varios algoritmos como: KNN, NB, SVM, J48 y RF, señalan que el mejor clasificador fue SVM + Polynomial Kernel.

Cheng et al. (2017) exponen en su estudio que para la detección de *trolling* utilizan el algoritmo LR el cual tuvo una precisión de 0,78 AUC.

J. Lee et al. (2018) para realizar la clasificación para la detección de trols, el estudio indica la utilización de un algoritmo basado en tensores llamado CTD descomposición tensorial compacta, este arroja un F1 score del 0,909 utilizando 10 tensores.

Borzemski, L. (2019) utilizan para la identificación del *trolling* algunos algoritmos como K-nearest Neighbors, NB Classifier, DT, LR, SVM, siendo el SVM el mejor puntuado según el score F1 72% y precisión 73%.

Para poder resumir en una tabla los estudios analizados y determinar cuál ha sido su fuente de datos primaria utilizada para el análisis, se crea una tabla que contiene las siguientes columnas: numeración, la cita de autor, el tema en español, los modelos y algoritmos utilizados, la fuente de datos utilizada, y el algoritmo que mejor resultado ha mostrado.

Cita de autor: Contiene la referencia a la cita del autor y de su estudio, a su vez esta cita está definida en la bibliografía

Tema: Se refiere al título del estudio traducido al español, en muchos casos se ha realizado un resumen

Modelos y algoritmos: Es la información referente a los modelos y algoritmos utilizados los estudios

Fuente: Se detallan las fuentes de información utilizadas en los diferentes estudios.

- Nombre de la red social
- Red social no identificada
- Fuente de datos independiente
- Redes de noticias

Resultado: Datos del algoritmo que mejor puntuación ha tenido

Tabla 38

Resumen de resultados.

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
1	(C. M. Lee et al., 2001)	Reconocimiento de emociones negativas	Linear discriminant classifier K-nearest neighborhood KNN	Fuente de datos independiente	No especificado
2	Ververidis et al. (2004)	Clasificación automática del habla emocional	Redes Bayesianas	Fuente de datos independiente	RB AUC 56%
3	Bishop (2013)	Abordaje del *trolling* mediante el desarrollo de una legislación	No abordados	Twitter Facebook	No especificado
4	Dlala et al. (2014)	Identificación de *trolls* en un marco incierto	Kmeans	Red social no definida	No especificado
5	Galán-García et al. (2014) Herrero et al. (2014)	Aprendizaje supervisado, para la detección de *trolls* en Twitter	Secuencial minimal optimation SMO Decision Trees Polly Kernel PK	Twitter	PK 96% AUC
6	De-La-Pena-Sordo et al. (2015)	Uso de modelos de compresión para filtrar comentarios Troll	KNN, Naïve Bayes, SVM, J48 y Random Forest	Redes de noticias	SVM + PK con el ACU 76.56%
7	Tsantarliotis et al. (2016)	Vulnerabilidades *troll* en redes sociales	Árboles de decisión	Reddit	No especificado
8	Mutlu et al. (2016)	Identificación de *trolls* y determinación de la conciencia del	KNN C4.5 Naïve Bayes	Twitter	C4.5 ACU 89%

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
		terror en redes sociales			
9	Chang et al. (2016)	Detección de rumores de usuarios extremos en Twitter	Algoritmo de clusterización no determinado	Twitter	No especificado
10	Authors (2017)	Detección de la agresión en publicaciones	SVM LR	Red social no definida	LR ACU 81.31%
11	Cheng et al. (2017)	Causas del comportamiento del troleo en línea	LR	Redes de noticias	LR ACU 78%
12	Fornacciari et al. (2018)	Sistema holístico para la detección del *trolling* en Twitter	Sequencial minial optimization SMO Naïve Bayes Random Forest	Twitter	SMO ACU 95%
13	Rasel et al. (2018)	Detección de comentarios ciber agresivos en redes sociales Redes	Decision Tree Random Forest SVM LR	Twitter	RF ACU 95%
14	Bellan et al. (2018)	Detección de comentarios inapropiados en noticas	Decision Tree Naïve Bayes LR	Twitter	LR F1 64%
15	Bastos et al. (2018)	Estudio de los *bots* y *trolls* que influyeron en la campaña del Brexit	No abordados	Twitter	No especificado
16	Bhatt et al. (2018)	Sistemas de detección de Trols	No abordados	Twitter	No especificado
17	Li et al. (2018)	Detección de miembros sospechosos en un servicio de apoyo	Random Forest SVM Gradient Boosted Tree	Fuente de datos independientes	RF ACU 98%

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
		emocional Online			
18	J. Lee et al. (2018)	Método rápido, preciso e interpretable para descomposiciones tensoriales estáticas y dinámicas	(CTD) descomposición tensorial compacta	Fuente de datos independientes	CDT F1 90%
19	Tian et al. (2019)	La cadena efectiva del troleo en las redes sociales	No abordados	Reddit Facebook Twitter Quora	No especificado
20	Kellner et al. (2019)	¿Qué es todo ese ruido?: Análisis y Detección de Propaganda en Twitter	Gradient Boosting Random Forest AdaBoost SVC LR Linear SVC y Knieghbors	Twitter	GB ACU 97%
21	(Luqyana et al., 2019)	Vecinos cercanos K como datos de equilibrio para la detección de *trolls*	KNN-Undersampling SVM	Twitter	KNN-U ACU 63.8%
22	Marouf et al. (2019)	Framework para detectar *trolls* en redes sociales utilizando psicolingüística	Multinomial Naïve Bayes Decision Tree Sequential Minimal Optimization	Facebook	MNB ACU 89.7%
23	Badawy et al. (2019)	¿Quién cae en la manipulación política en línea?	Gradient Boosting	Twitter	GB AUC 96%
24	Borzemski, L. (2019)	Recognition of Antisocial Behavior in Online Discussions	K-nearest Neighbors Naïve BayesClassifier Decision Trees	Redes de noticias	SVM AUC 73%

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
			LR SVM		
25	Lai et al. (2019)	Polaridad de postura en debates políticos	SVM	Twitter	No especificado
26	Hitkul et al. (2019)	Detección de las imágenes propensas al *trolling* en Instagram	VGG16 VGG19 en conjuntos con SVM Inception V2 y V3, RestNet50	Instagram	VGG+VGG19 +SVM ACU 61.8%
27	Im et al. (2020)	Modelar e identificar cuentas de *trolls* rusos en Twitter	LR Decision Tree Adaptive Boosted	Twitter	ADT AUC 99%
28	Dewhurst et al. (2020)	Dinámicas no cooperativas en la interferencia electoral	No-U-Turn Sampler Gaussian Process Bayesian	Twitter	No especificado
29	Islam et al. (2020)	Detección de Ciberacoso en redes sociales utilizando enfoques de aprendizaje automático	Support Vector Machine Random Forest Decision Tree Naïve Bayes	Facebook Twitter	SVM AUC 81%
30	Sarsam et al. (2020)	Detección de sarcasmo usando un grupo de algoritmos de aprendizaje con datos de Twitter	Support Vector Machine LR Naïve Bayes Random Forest	Red social no definida	SVM AUC 91.8%
31	Lundberg et al. (2020)	Los *trolls* de Twitter	J48-HP	Twitter	J48-HP AUC 99%
32	Gamova et al. (2020)	Detección de comentarios falsos y	Random Forest LR GaussianNB	Twitter	No especificado

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
		provocativos en las redes sociales usando aprendizaje automático	Linear SVC		
33	Alhazbi S. (2020)	Enfoques de aprendizaje automático basados en el comportamiento para identificar *trolls* patrocinados por el estado en Twitter	Decision Tree Random Forest Adaboost Gradient Boost	Twitter	GB AUC 94%
34	Ramiandris (2020)	Identificación de agresión en publicaciones: dos enfoques de aprendizaje automático	Random Forest Regression Linear Convolutional neural network CNN LSTM	Facebook	RF+RL F1 58%
35	Modha et al. (2020)	Identificación de agresión en publicaciones	SVM LR Convolutional Neural Network CNN	Facebook	CNN F1 64%
36	Singh et al. (2020)	Clasificación Multi Nivel de los comentarios negativos	OneVsAll Logistic Multi-Level Bi-LSTM y Bi-Gru	Fuentes de datos independientes	Bi-LSTM + Bi-GRU ACU 97.8%
37	Smitha et al. (2021)	Detección automática de actores influyentes en redes de desinformación	Random Forest Extra-Trees ET xgBoost	Twitter	No especificado
38	Sadiq et al. (2021)	Detección de agresiones mediante modelos de redes	CNN-LSTM CNN-BiLSTM Multi-Layer Perceptron MLP TF-IDF	Twitter	TF-IDF AUC 90%

No.	Cita	Tema	Algoritmos	Fuente de datos utilizada	Resultado
		neuronales profundas en Twitter			
39	Jachim et al. (2021)	Detección en tiempo real de narrativas de *trolling* en Twitter durante las elecciones estadounidenses de 2020	No abordados	Twitter	No especificado
40	MacHova et al. (2021)	Algoritmos de machine learning en el reconocimiento de *trolls*	SVM Multinomial Naïve Bayes LR	Facebook	MNB AUC 88%

Nota. Resumen de modelos y algoritmos utilizados para la detección del trolling.

3.6 Resumen de las aplicaciones desarrolladas

En la Tabla 42 se resume las caracteristicas de los sistemas informáticos que hasta el momento se han desarrollado, la fuente de datos y sus algoritmos utilizados.

Tabla 39

Resumen de resultados

No.	Cita	Nombre del software	Algoritmos	Fuente de datos utilizada
1	(Bhatt et al., 2018)	SMC4	La aplicación, llamada SMC4 Lite, afirma ser el "primer software anti-troleo del mundo". Este software, captura y analiza todos los mensajes entrantes y salientes de las redes sociales de la cuenta de Twitter y bloquea automáticamente	Twitter

No.	Cita	Nombre del software	Algoritmos	Fuente de datos utilizada
			cualquier intento de acoso, blasfemia, sexismo, racismo y otro lenguaje que causa conflicto.	
2	(Li et al., 2018)	TeaFilter	Detecta los comportamientos sospechosos	Bases de datos independientes
3	Badawy et al. (2019)	Botometer	Analiza la actividad de una cuenta de Twitter y le da una puntuación. Las puntuaciones más altas significan más actividad similar a la de un *bot*, pudiendo identificar los automatismos que generan ciberacoso o *trolling*	Twitter
4	(Modha et al., 2020)	Ad-on	Se usa como complemento para los navegadores que ayudaría al control del *trolling* o acoso y agresiones fuertes el estudio ha utilizado varios clasificadores como Support Vector Machine SVM.	Red social no definida
5	(Jachim et al., 2021)	TrollHunter2000	Es una aplicación para detectar el *trolling* en tiempo real en Twitter.	Twitter

Nota. Resumen de los sistemas creados para la detección del trolling.

CAPÍTULO 4

CONCLUSIONES Y TRABAJOS FUTUROS

El objetivo principal de este trabajo es el análisis sistemático de los estudios realizados para detectar el *trolling* o el ciberacoso, fue realizado mediante un análisis cronológico y profundo al contenido de los documentos realizados por diferentes estudiosos del tema que han utilizado técnicas de minería de datos y aprendizaje automático, permitiendo así disponer de valiosos resultados en la detección y prevención de esta realidad presente en los sitios web, foros, y en mayor grado en las redes sociales.

Otro objetivo importante y como se ha descrito a lo largo de este trabajo es determinar cuáles fueron las fuentes de información utilizadas en los estudios. Se identificó que la mayoría de los datos para los diferentes análisis han sido provenientes de la red social Twitter (26) y muy pocas veces Facebook (7), Reddit (2), Instagram (1), Quora (1), Independientes (9), esto quizá sea un problema ya que puede existir resultados diferentes al no incluir en estos análisis otras fuentes de información no consideradas, lo que quizá podría significar un sesgo.

Para la realización de este estudio fue necesario desarrollar una metodología basada en diferentes trabajos y métodos sugeridos sobre revisiones sistemáticas que permita recopilar, clasificar y analizar todos los trabajos de minería de datos y aprendizaje automático pudiendo determinar cuáles han sido los resultados obtenidos en relación con los algoritmos y modelos empleados. La identificación del grupo de algoritmos más usados fue un objetivo más que se pudo identificar en una de las tablas resumen de este estudio, pero no es determinante como para poder sugerir un algoritmo que entregue una alta efectividad en su resultado. También se pudo evidenciar que en algunos estudios fue complicado obtener los datos de las redes sociales para sus análisis, ya que estos no son datos abiertos para consumo de cualquier persona.

Las leyes de protección de datos que se encuentran vigentes en los países prohíben a las empresas a divulgar información lo que probablemente sea una causa por la cual empresas como Facebook o Instagram no entregan información fácilmente, o quizá los motivos sean más comerciales, de todas maneras, las empresas que ofrecen servicios de *social netwoking* deberían ofrecer entre sus prestaciones mecanismos para erradicar el *trolling* en un nivel aceptable.

Ya en el análisis de los estudios recopilados se logró identificar los algoritmos más utilizados, unos con mejores resultados en su precisión sobre la clasificación del *trolling* como lo son SVM 91%, RF 98%, GB 97%, ADT 99%. Adicionalmente se pudo observar muchos algoritmos usados que ya son conocidos como Random Forest, Regression Lineal, Naïve Bayes, Kmeans, SVM, SMO, Gradient Boost, CNN, estos son los que más frecuentemente se encontró; también se nombran algunos que quizá son poco conocidos pero que han entregado valores de clasificación interesantes como Polly Kernel 96% AUC, CTD F1 90%, y otros que también fueron usados pero no registran valores de resultado, estos son No-U-Turn Sampler y Gaussian Process.

Como se ha mencionado anteriormente ningún modelo entrega un 100% de certeza en la clasificación o detección del contenido que identifica al *trolling*, quizá esto se deba a la poca información que entregan las empresas que son dueñas de los servicios de redes sociales y también al contenido semántico (lenguaje natural) que cada país tiene o cada cultura expresa, ya que las palabras y los conceptos varían de acuerdo a las circunstancias; por tanto los modelos de entrenamientos deben ampliarse tomando en cuenta también datos de varios idiomas, culturas y jergas posibles.

Se debe puntualizar que en este momento ya se dispone de aplicaciones que han sido desarrolladas para la identificación del *trolling* en redes sociales (5), seguramente estas seguirán un ciclo de vida que permita una mejora continua de sus procesos de detección. Los estudios revisados en los que se desarrollaron sistemas de software para la detección del *trolling* marcan un punto de referencia para otros investigadores o empresas que requieran una base científica para construir servicios de detección de patrones de contenido ofensivo en el lenguaje natural.

Algo que es notable en algunos estudios analizados (4) es la importancia que se da a la identificación del sesgo político que podría causar el *trolling* ya que mediante esto se puede influenciar en el pensamiento político en algunas personas, este tema se debe abordar con mayor profundidad en otra investigación, no obstante, en muchos estudios descritos en este trabajo se evidencia un porcentaje importante de los estados o países que han puesto recursos para promover la generación de *trolling* influyente en grupos humanos.

Se identifica que las tecnologías de la información han evolucionado ágilmente entregando herramientas de desarrollo ágil y de almacenaje de datos más robustos sobre los cuales se puede almacenar y extraer más información, esto abrió la puerta a que las redes sociales digitales se

desarrollen rápidamente pudiendo incorporar servicios que permiten a los usuarios compartir video, imágenes, texto siendo este último el medio más usado para la generación del *trolling*. Por lo tanto, los mecanismos de detección deben extenderse no solo al análisis de texto sino también al video, al audio y a las imágenes algo que se pudo observar mínimamente (1) en los diferentes estudios. Solamente en un estudio abordaba el análisis de imágenes, pero desde la perspectiva de causa efecto.

También se sugiere a las empresas y estados que son dueñas de la información dotar de mecanismos legales y técnicos para el uso de sus datos en fines científicos ya que actualmente es limitado el acceso a su información reduciendo así la posibilidad de desarrollar mecanismos de control del *trolling.*

La clasificación de los estudios realizados permite identificar que la mayor parte de ellos se centran en el reconocimiento del *trolling* mediante el análisis e identificación de texto (38) mientras que solamente se identifica que un (1) estudio analiza esta realidad desde la distinción de las imágenes que causan el *trolling,* este resultado aporta con un nivel de conocimiento para que los investigadores requieran enfocar sus futuros análisis en las técnicas que no se han desarrollado aún.

Adicionalmente este trabajo ofrece un marco de referencia ágil de 5 actividades que contribuyen al análisis de las revisiones sistemáticas relacionadas al tema propuesto simplificando así el abordaje y la obtención del conocimiento para las investigaciones futuras.

Finalmente se concluye que mediante este estudio se entrega una visión consolidada de las investigaciones relacionadas a la detección y mitigación del *trolling* proporcionando un punto de partida mediante el conocimiento de las capacidades desarrolladas hasta esta esta fecha, las cuales serán se utilidad para las personas que requieran desarrollar a futuro técnicas para la detección de esta realidad.

BIBLIOGRAFÍA

Abigail Orús. (2022). *Evolución del número de suscriptores de Netflix en el mundo del tercer trimestre de 2011 hasta el cuarto trimestre de 2021*. 1. Retrieved from https://www.citethisforme.com/languages/es/apa/como-citar-una-pagina-web#:~:text=El formato para referenciar una,Nombre de la página web.

Authors, F. (2017). *The Dark Side of News Community Forums : Opinion Manipulation Trolls Article information : To cite this document : About Emerald www.emeraldinsight.com*.

Badawy, A., Lerman, K., & Ferrara, E. (2019). Who falls for online political manipulation? *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 162–168. doi: 10.1145/3308560.3316494

Basalingappa, A., Subhas, M. S., & Tapariya, R. (2019). Understanding Likes on Facebook: An Exploratory Study. *Online Journal of Communication and Media Technologies*, *6*(3). doi: 10.29333/ojcmt/2566

Bastos, M., & Mercea, D. (2018). The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). doi: 10.1098/rsta.2018.0003

Bellan, P., & Strapparava, C. (2018). Detecting Inappropriate Comments to News. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11298 LNAI. Springer International Publishing. doi: 10.1007/978-3-030-03840-3_30

Bhatt, U., Iyyani, D., Jani, K., & Mali, S. (2018). Troll-Detection Systems Limitations of Troll Detection Systems and AI/ML Anti-Trolling Solution. *2018 3rd International Conference for Convergence in Technology, I2CT 2018*, 1–6. doi: 10.1109/I2CT.2018.8529342

Bishop, J. (2013). The art of trolling law enforcement: a review and model for implementing "flame trolling" legislation enacted in Great Britain (1981-2012). *International Review of Law, Computers and Technology*, *27*(3), 301–318. doi: 10.1080/13600869.2013.796706

Ceballos Saavedra, M. (2014). Los origenes de las redes sociales y los medios de comunicacion. *Pensamiento y Cultura*, *11*(2), 263–275.

Chang, C., Zhang, Y., Szabo, C., & Sheng, Q. Z. (2016). Extreme user and political rumor detection on twitter. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10086 LNAI*, 751–763. doi: 10.1007/978-3-319-49586-6_54

Cheng, J., Bernstein, M., Danescu-Niculescu-mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 1217–1230. doi: 10.1145/2998181.2998213

Dasilva, J. P., Arratibel, A. G., Aierdi, K. M., Galdospín, T. M., Castillo, I. M., Mardaras, L. I., González, M. R., & Santamarina, D. R. (2013). Las empresas en Facebook y Twitter . Situación actual y estrategias comunicativas Companies on Facebook and Twitter . Current situation and. *Revista Latina de Comunicación Social*, *68*, 676–695. doi: 10.4185/RLCS-

2013-996

De-La-Pena-Sordo, J., Pastor-Lopez, I., Santos, I., & Bringas, P. G. (2015). Using compression models for filtering troll comments. *Proceedings of the 2015 10th IEEE Conference on Industrial Electronics and Applications, ICIEA 2015*, 655–660. doi: 10.1109/ICIEA.2015.7334191

Dewhurst, D. R., Danforth, C. M., & Dodds, P. S. (2020). Noncooperative dynamics in election interference. *Physical Review E*, *101*(2), 1–25. doi: 10.1103/PhysRevE.101.022307

Dlala, I. O., Attiaoui, D., Martin, A., & Yaghlane, B. Ben. (2014). Trolls Identification within an Uncertain Framework. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2014-Decem*, 1011–1015. doi: 10.1109/ICTAI.2014.153

FECYT. (2017a). *Guía Metodológica Para La Creación De Una Clasificación De Revistas En Ciencias Humanas Y Sociales, Destinada a Las Agencias De Evaluación Del Mérito Docente E Investigador*. 41. Retrieved from https://calidadrevistas.fecyt.es/sites/default/files/informes/guia_ccss_hum_def.pdf

FECYT, F. E. para la C. y T. (2017b). *Funcionalidades avanzadas en Scopus*. 53. Retrieved from https://www.recursoscientificos.fecyt.es/sites/default/files/2018_04_03_scopus_basico_avanzado_abril_18.pdf

Fornacciari, P., Mordonini, M., Poggi, A., Sani, L., & Tomaiuolo, M. (2018). A holistic system for troll detection on Twitter. *Computers in Human Behavior*, *89*, 258–268. doi: 10.1016/j.chb.2018.08.008

Galán-García, P., De La Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2014). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, *24*(1), 42–53. doi: 10.1093/jigpal/jzv048

Gamova, A. A., Horoshiy, A. A., & Ivanenko, V. G. (2020). Detection of Fake and Provokative Comments in Social Network Using Machine Learning. *Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus 2020*, 309–311. doi: 10.1109/EIConRus49466.2020.9039423

Ganguli, R., Mehta, A., & Sen, S. (2020). A Survey on Machine Learning Methodologies in Social Network Analysis. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 484–489. doi: 10.1109/ICRITO48877.2020.9197984

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., … Wu, D. M. (2017). A large human-labeled corpus for online harassment research. *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*, 229–233. doi: 10.1145/3091478.3091509

Guidelines, V. I. (2007). Enrich Your Experience. *Online*.

Herrera, R. S. (2021). Trolling in the framework of political practices. *Contratexto*, *35*, 91–111. doi: 10.26439/CONTRATEXTO2021.N035.5034

Herrero, Á., Baruque, B., Klett, F., Abraham, A., Snášel, V., De Carvalho, A. C. P. L. F., Bringas, P. G., Zelinka, I., Quintián, H., & Corchado, E. (2014). Preface. *Advances in Intelligent Systems and Computing*, *239*, v–vi. doi: 10.1007/978-3-319-01854-6

Hitkul, H., Shah, R. R., Kumaraguru, P., & Satoh, S. (2019). Maybe look closer? Detecting trolling prone images on Instagram. *Proceedings - 2019 IEEE 5th International Conference on Multimedia Big Data, BigMM 2019*, 448–456. doi: 10.1109/BigMM.2019.00029

Hofmann, P., Samp, C., & Urbach, N. (2020). Robotic process automation. *Electronic Markets*, *30*(1), 99–106. doi: 10.1007/s12525-019-00365-8

Im, J., Chandrasekharan, E., Sargent, J., Lighthammer, P., Denby, T., Bhargava, A., Hemphill, L., Jurgens, D., & Gilbert, E. (2020). Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 1–10. doi: 10.1145/3394231.3397889

Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying Detection on Social Networks Using Machine Learning Approaches. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*, *April*. doi: 10.1109/CSDE50874.2020.9411601

Jachim, P., Sharevski, F., & Pieroni, E. (2021). TrollHunter2020: Real-time detection of trolling narratives on twitter during the 2020 u.s. elections. *IWSPA 2021 - Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics*, 55–65. doi: 10.1145/3445970.3451158

Jack Dorsey, Noah Glass, Biz Stone, E. W. (2006). *Twitter*.

Kashish Hora. (n.d.). *Exist, This person doesn´t*. Retrieved from https://thispersondoesnotexist.com/

Kellner, A., Rangosch, L., Wressnegger, C., & Rieck, K. (2019). *Political Elections Under (Social) Fire? Analysis and Detection of Propaganda on Twitter*. 1–20. Retrieved from http://arxiv.org/abs/1912.04143

Kevin Systrom, Mike Krieger (Burbn, I. . (2010). *Instagram*.

Lai, M., Tambuscio, M., Patti, V., Ruffo, G., & Rosso, P. (2019). Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter. *Data and Knowledge Engineering*, *124*(August), 101738. doi: 10.1016/j.datak.2019.101738

Lee, C. M., Narayanan, S., & Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. *2001 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2001 - Conference Proceedings*, 240–243. doi: 10.1109/ASRU.2001.1034632

Lee, J., Choi, D., & Sael, L. (2018). Ctd: Fast, accurate, and interpretable method for static and dynamic tensor decompositions. *PLoS ONE*, *13*(7), 1–18. doi: 10.1371/journal.pone.0200579

Li, Y., Kim, D. W., Zhang, J., & Doran, D. (2018). Detecting suspicious members in an online emotional support service. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *255*, 22–42. doi: 10.1007/978-3-030-01704-0_2

Lundberg, J., & Laitinen, M. (2020). Twitter trolls: a linguistic profile of anti-democratic discourse. *Language Sciences*, *79*(xxxx), 101268. doi: 10.1016/j.langsci.2019.101268

Lupano Perugini, M. L., & Castro Solano, A. (2021). Rasgos de personalidad, bienestar y malestar psicológico en usuarios de redes sociales que presentan conductas disruptivas online. *Interdisciplinaria Revista de Psicología y Ciencias Afines*, *38*(2), 7–23. doi: 10.16888/interd.2021.38.2.1

Luqyana, W. A., Ahmadie, B. L., & Supianto, A. A. (2019). K-Nearest Neighbors Undersampling as Balancing Data for Cyber Troll Detection. *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, 322–325. doi: 10.1109/SIET48054.2019.8986079

MacHova, K., Porezany, M., & Hreskova, M. (2021). Algorithms of Machine Learning in Recognition of Trolls in Online Space. *SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 349–353. doi: 10.1109/SAMI50585.2021.9378699

Mark Zuckerberg. (2004). *Facebook*.

Marouf, A. Al, Ajwad, R., & Ashrafi, A. F. (2019). Looking behind the Mask: A framework for Detecting Character Assassination via Troll Comments on Social media using Psycholinguistic Tools. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–5. doi: 10.1109/ECACE.2019.8679154

Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance. *Expert Systems with Applications*, *161*, 113725. doi: 10.1016/j.eswa.2020.113725

Moreno, A., Armengol, E., Bejar, E., Belanche, L., Cortez, U., Gavalda, R., Lopez, B., & Sanchez, M. (2004). *Aprendizaje automático*.

Mutlu, B., Mutlu, M., Oztoprak, K., & Dogdu, E. (2016). Identifying trolls and determining terror awareness level in social networks using a scalable framework. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 1792–1798. doi: 10.1109/BigData.2016.7840796

Orallo, J. H., Quintana, M. J. R., & Ramirez, C. F. (2004). *Introducción a la minería de datos*.

Paperpile. (2021). *The top list academic research databases*. Retrieved from https://paperpile.com/g/academic-research-databases/

Ramiandrisoa, F. (2020). Aggression identification in posts - Two machine learning approaches. *CEUR Workshop Proceedings*, *2606*, 40–49.

Rasel, R. I., Sultana, N., Akhter, S., & Meesad, P. (2018). *Detection of Cyber-Aggressive Comments on Social Media Networks*. 37–41. doi: 10.1145/3278293.3278303

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2021). Aggression detection through deep neural model on Twitter. *Future Generation Computer Systems*, *114*, 120–129. doi: 10.1016/j.future.2020.07.050

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, *62*(5), 578–598. doi: 10.1177/1470785320921779

Singh, J., & Nongmeikapam, K. (2020). Negative Comments Multi-Label Classification. *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, 379–

385. doi: 10.1109/ComPE49325.2020.9200131

Smitha, S. T., Kaoa, E. K., MacKina, E. D., Shaha, D. C., Simeka, O., & Rubin, D. B. (2021). Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(4). doi: 10.1073/PNAS.2011216118

Statista. (2022). *Índice de penetración de Internet en el mundo desde 2009 hasta 2019, por región*. 1. Retrieved from https://es.statista.com/estadisticas/635464/indice-de-penetracion-de-internet-en-el-mundo-2009--por-region/

Tebes, G., Peppino, D., Becker, P., & Olsina, L. (2019). Especificación del modelo de proceso para una revisión sistemática de literatura. *XXII Ibero-American Conference on Software Engineering, CIbSE 2019*, *April*, 391–404.

Tian, H., & Chen, P. Y. (2019). "I'm in the center of the vortex": The affective chain of social media trolling. *Proceedings of the Association for Information Science and Technology*, *56*(1), 778–779. doi: 10.1002/pra2.173

Tsantarliotis, P., Pitoura, E., & Tsaparas, P. (2016). Troll vulnerability in online social networks. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 1394–1396. doi: 10.1109/ASONAM.2016.7752428

Uribe-saavedra, F. (2011). *Redes sociales digitales como herramienta de marketing - Un estudio de casos CONGRESO NACIONAL DE ACEDE SEPTIEMBRE 2011 , BARCELONA MARKETING – UN ESTUDIO DE CASOS Felipe Uribe Saavedra 1 Universitat Autònoma de Barcelona Josep Rialp Criado Universitat Au*. *July 2015*.

Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*. doi: 10.1109/icassp.2004.1326055

ANEXOS

A continuación, se muestra la Tabla 40 obtenida de la validación de cada uno de los estudios en el ranking de revistas científicas.

Tabla 40

Clasificación de estudios

No.	Autor	Título	Año	Origen	Revista
1	Sahni A., Raja N.	Analyzation and detection of cyberbullying: A Twitter based Indian case study	2018	Scopus	Q4
2	Bellan P., Strapparava C.	Detecting Inappropriate Comments to News	2018	Scopus	Q3
3	Chang C., Zhang Y., Szabo C., Sheng Q.Z.	Extreme user and political rumor detection on twitter	2016	Scopus	Q3
4	Lai M., Tambuscio M., Patti V., Ruffo G., Rosso P.	Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter	2019	Scopus	Q2
5	Bishop J.	The art of trolling law enforcement: a review and model for implementing 'flame trolling' legislation enacted in Great Britain (1981-2012)	2013	Scopus	Q2

No.	Autor	Título	Año	Origen	Revista
6	Sarsam S.M., Al-Samarraie H., Alzahrani A.I., Wright B.	Sarcasm detection using machine learning algorithms in Twitter: A systematic review	2020	Scopus	Q2
7	Smitha S.T., Kaoa E.K., MacKina E.D., Shaha D.C., Simeka O., Rubin D.B.	Automatic detection of influential actors in disinformation networks	2021	Scopus	Q1
8	Sadiq S., Mehmood A., Ullah S., Ahmad M., Choi G.S., On B.-W.	Aggression detection through deep neural model on Twitter	2021	Scopus	Q1
9	Lundberg J., Laitinen M.	Twitter trolls: a linguistic profile of anti-democratic discourse	2020	Scopus	Q1
10	Dewhurst D.R., Danforth C.M., Dodds P.S.	Noncooperative dynamics in election interference	2020	Scopus	Q1
11	Bail C.A., Guay B., Maloney E., Combs A., Sunshine Hillygus D., Merhout F., Freelon D., Volfovsky A.	Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017	2020	Scopus	Q1
12	Alhazbi S.	Behavior-based machine learning approaches to identify state-sponsored trolls on twitter	2020	Scopus	Q1

No.	Autor	Título	Año	Origen	Revista
13	Fornacciari P., Mordonini M., Poggi A., Sani L., Tomaiuolo M.	A holistic system for troll detection on Twitter	2018	Scopus	Q1
14	Lee J., Choi D., Sael L.	Ctd: Fast, accurate, and interpretable method for static and dynamic tensor decompositions	2018	Scopus	Q1
15	Bastos M., Mercea D.	The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign	2018	Scopus	Q1
16	Tsantarliotis P., Pitoura E., Tsaparas P.	Defining and predicting troll vulnerability in online social media	2017	Scopus	Q1
17	Galán-García P., De La Puerta J.G., Gómez C.L., Santos I., Bringas P.G.	Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying	2014	Scopus	Q1
18	Modha S., Majumder P., Mandl T., Mandalia C.	Detecting and visualizing hate speech in social media: A cyber-Watchdog for surveillance	2020	Scopus	Q1
19	Ganguli R., Mehta A., Sen S.	A Survey on Machine Learning Methodologies in Social Network Analysis	2020	Scopus	ND

No.	Autor	Título	Año	Origen	Revista
20	Jachim P., Sharevski F., Pieroni E.	TrollHunter2020: Real-time detection of trolling narratives on twitter during the 2020 u.s. elections	2021	Scopus	N/D
21	MacHova K., Porezany M., Hreskova M.	Algorithms of Machine Learning in Recognition of Trolls in Online Space	2021	Scopus	N/D
22	Islam M.M., Uddin M.A., Islam L., Akter A., Sharmin S., Acharjee U.K.	Cyberbullying Detection on Social Networks Using Machine Learning Approaches	2020	Scopus	N/D
23	Im J., Chandrasekharan E., Sargent J., Lighthammer P., Denby T., Bhargava A., Hemphill L., Jurgens D., Gilbert E.	Still out there: Modeling and Identifying Russian Troll Accounts on Twitter	2020	Scopus	N/D
24	Singh J., Nongmeikapam K.	Negative Comments Multi-Label Classification	2020	Scopus	N/D
25	Kellner A., Wressnegger C., Rieck K.	What's all that noise: Analysis and detection of propaganda on Twitter	2020	Scopus	N/D
26	Gamova A.A., Horoshiy A.A., Ivanenko V.G.	Detection of Fake and Provokative Comments in Social Network Using Machine Learning	2020	Scopus	N/D

No.	Autor	Título	Año	Origen	Revista
27	Machová K., Kolesár D.	Recognition of Antisocial Behavior in Online Discussions	2020	Scopus	N/D
28	Ramiandrisoa F.	Aggression identification in posts - Two machine learning approaches	2020	Scopus	N/D
29	Luqyana W.A., Ahmadie B.L., Supianto A.A.	K-Nearest Neighbors Undersampling as Balancing Data for Cyber Troll Detection	2019	Scopus	N/D
30	Hitkul H., Shah R.R., Kumaraguru P., Satoh S.	Maybe look closer? Detecting trolling prone images on Instagram	2019	Scopus	N/D
31	Badawy A., Lerman K., Ferrara E.	Who falls for online political manipulation?	2019	Scopus	N/D
32	Marouf A.A., Ajwad R., Ashrafi A.F.	Looking behind the Mask: A framework for Detecting Character Assassination via Troll Comments on social media using Psycholinguistic Tools	2019	Scopus	N/D
33	Tian H., Chen P.-Y.	“I'm in the center of the vortex”: The affective chain of social media trolling	2019	Scopus	N/D

No.	Autor	Título	Año	Origen	Revista
34	Bhatt U., Iyyani D., Jani K., Mali S.	Troll-Detection Systems Limitations of Troll Detection Systems and AI/ML Anti-Trolling Solution	2018	Scopus	N/D
35	Mihaylov T., Mihaylova T., Nakov P., Màrquez L., Georgiev G.D., Koychev I.K.	The dark side of news community forums: opinion manipulation trolls	2018	Scopus	N/D
36	Rasel R.I., Sultana N., Akhter S., Meesad P.	Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach	2018	Scopus	N/D
37	Li Y., Kim D.W., Zhang J., Doran D.	Detecting suspicious members in an online emotional support service	2018	Scopus	N/D
38	Golbeck J., Ashktorab Z., Banjo R.O., Berlinger A., Bhagwan S., Buntain C., Cheakalos P., Geller A.A., Gergory Q., Gnanasekaran R.K., Gunasekaran R.R., Hoffman K.M., Hottle J., Jienjitlert V., Khare S., Lau R., Martindale M.J., Naik S., Nixon H.L., Ramachandran P., Rogers K.M., Rogers L., Sarin	A large human-labeled corpus for online harassment research	2017	Scopus	N/D

No.	Autor	Título	Año	Origen	Revista
	M.S., Shahane G., Thanki J., Vengataraman P., Wan Z., Wu D.M.				
39	Cheng J., Bernstein M., Danescu-Niculescu-mizil C., Leskovec J.	Anyone can become a troll: Causes of trolling behavior in online discussions	2017	Scopus	N/D
40	Tsantarliotis P., Pitoura E., Tsaparas P.	Troll vulnerability in online social networks	2016	Scopus	N/D
41	Wu Z., Aggarwal C.C., Sun J.	The troll-trust model for ranking in signed networks	2016	Scopus	N/D
42	Mutlu B., Mutlu M., Oztoprak K., Dogdu E.	Identifying trolls and determining terror awareness level in social networks using a scalable framework	2016	Scopus	N/D
43	Dlala I.O., Attiaoui D., Martin A., Yaghlane B.B.	Trolls Identification within an Uncertain Framework	2014	Scopus	N/D

Nota. Clasificación de los estudios por año y por revista