Universidad Continental

Escuela de Posgrado

MAESTRÍA EN CIENCIA DE DATOS

Tesis

Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos

Joel Jesus Bastidas Valdivia
Giorgio Giacomo Crose Guzman
Luis antony Lopez Quiroz

Para optar el Grado Académico de
Maestro en Ciencia en Ciencia de atos

Huancayo, 2025

Universidad Continental

ANEXO 6

INFORME DE CONFORMIDAD DE ORIGINALIDAD DEL

TRABAJO DE INVESTIGACIÓN

A	:	Mg. Jaime Sobrados Tapia Director Académico de la Escuela de Posgrado
DE	:	**Kevin Rafael Palomino Pacheco** Asesor del Trabajo de Investigación
ASUNTO	:	Remito resultado de evaluación de originalidad de Trabajo de Investigación
FECHA	:	15 de febrero del 2025

Con sumo agrado me dirijo a vuestro despacho para saludarlo y en vista de haber sido designado Asesor del Trabajo de Investigación/Tesis/Artículo Científico titulado "**Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos**", perteneciente a los **Bach. Joel Jesus Bastidas Valdivia**, **Bach. Giorgio Giacomo Crose Guzman**, y **Bach. Luis Antony Lopez Quiroz**, de la **Maestría en Ciencia de Datos**; se procedió con el análisis del documento mediante la herramienta "Turnitin" y se realizó la verificación completa de las coincidencias resaltadas por el software, cuyo resultado es **7%** de similitud (informe adjunto) sin encontrarse hallazgos relacionados con plagio. Se utilizaron los siguientes filtros:

- Filtro de exclusión de bibliografía — SÍ [x] NO [ ]
- Filtro de exclusión de grupos de palabras menores (Máximo nº de palabras excluidas: < **40**) — SÍ [x] NO [ ]
- Exclusión de fuente por trabajo anterior del mismo estudiante — SÍ [x] NO [ ]

En consecuencia, se determina que el trabajo de investigación constituye un documento original al presentar similitud de otros autores (citas) por debajo del porcentaje establecido por la Universidad.

Recae toda responsabilidad del contenido de la tesis sobre el autor y asesor, en concordancia a los principios de legalidad, presunción de veracidad y simplicidad, expresados en el Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI y en la Directiva 003-2016-R/UC.

Esperando la atención a la presente, me despido sin otro particular y sea propicia la ocasión para renovar las muestras de mi especial consideración.

Atentamente,

Kevin Rafael Palomino Pacheco
DNI: 1045711819

Universidad Continental

DECLARACIÓN JURADA DE AUTENTICIDAD

Yo, JOEL JESUS BASTIDAS VALDIVIA, identificado con Documento Nacional de Identidad N° 40024611, egresado de la MAESTRÍA EN CIENCIA DE DATOS, de la Escuela de Posgrado de la Universidad Continental, declaro bajo juramento lo siguiente:

1. La Tesis titulada "*Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos*", es de mi autoría, la misma que presento para optar el Grado Académico de MAESTRO EN CIENCIA DE DATOS.
2. La Tesis no ha sido plagiada ni total ni parcialmente, para lo cual se han respetado las normas internacionales de citas y referencias para las fuentes consultadas, por lo que no atenta contra derechos de terceros.
3. La Tesis es original e inédita, y no ha sido realizada, desarrollada o publicada, parcial ni totalmente, por terceras personas naturales o jurídicas. No incurre en autoplagio; es decir, no fue publicada ni presentada de manera previa para conseguir algún grado académico o título profesional.
4. Los datos presentados en los resultados son reales, pues no son falsos, duplicados, ni copiados, por consiguiente, constituyen un aporte significativo para la realidad estudiada.

De identificarse fraude, falsificación de datos, plagio, información sin cita de autores, uso ilegal de información ajena, asumo las consecuencias y sanciones que de mi acción se deriven, sometiéndome a las acciones legales pertinentes.

Lima, 28 de noviembre de 2024.

JOEL JESUS BASTIDAS VALDIVIA
DNI. N° 40024611

Huella

Arequipa
Av. Los Incas S/N,
José Luis Bustamante y Rivero
(054) 412 030

Calle Alfonso Ugarte 607, Yanahuara
(054) 412 030

Huancayo
Av. San Carlos 1980
(064) 481 430

Cusco
Urb. Manuel Prado - Lote B, N° 7 Av. Collasuyo
(084) 480 070

Sector Angostura KM. 10,
carretera San Jerónimo - Saylla
(084) 480 070

Lima
Av. Alfredo Mendiola 5210, Los Olivos
(01) 213 2760

Jr. Junín 355, Miraflores
(01) 213 2760

DECLARACIÓN JURADA DE AUTENTICIDAD

Yo, GIORGIO GIACOMO CROSE GUZMAN, identificado con Documento Nacional de Identidad N° 70126005, egresado de la MAESTRÍA EN CIENCIA DE DATOS, de la Escuela de Posgrado de la Universidad Continental, declaro bajo juramento lo siguiente:

1. La Tesis titulada "*Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos*", es de mi autoría, la misma que presento para optar el Grado Académico de MAESTRO EN CIENCIA DE DATOS.
2. La Tesis no ha sido plagiada ni total ni parcialmente, para lo cual se han respetado las normas internacionales de citas y referencias para las fuentes consultadas, por lo que no atenta contra derechos de terceros.
3. La Tesis es original e inédita, y no ha sido realizada, desarrollada o publicada, parcial ni totalmente, por terceras personas naturales o jurídicas. No incurre en autoplagio; es decir, no fue publicada ni presentada de manera previa para conseguir algún grado académico o título profesional.
4. Los datos presentados en los resultados son reales, pues no son falsos, duplicados, ni copiados, por consiguiente, constituyen un aporte significativo para la realidad estudiada.

De identificarse fraude, falsificación de datos, plagio, información sin cita de autores, uso ilegal de información ajena, asumo las consecuencias y sanciones que de mi acción se deriven, sometiéndome a las acciones legales pertinentes.

Lima, 28 de noviembre de 2024.

GIORGIO GIACOMO CROSE GUZMAN
DNI. N° 70126005

Huella

DETECCIÓN DE SEÑALES DE VIOLENCIA EN REDES SOCIALES: UN ENFOQUE DESDE LA CIENCIA DE DATOS

INFORME DE ORIGINALIDAD

7%	7%	3%	2%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	1%
2	Submitted to Universidad Continental Trabajo del estudiante	1%
3	uvadoc.uva.es Fuente de Internet	1%
4	repositorio.continental.edu.pe Fuente de Internet	1%
5	sedici.unlp.edu.ar Fuente de Internet	1%
6	www.johncaicedo.com.co Fuente de Internet	<1%
7	rua.ua.es Fuente de Internet	<1%
8	ia.university Fuente de Internet	<1%
	Submitted to imfice	

9	Trabajo del estudiante	<1 %
10	Fabiola Montero, Nelson Montilla, Julio Arcia. "Algoritmos de aprendizaje automático en la predicción del rendimiento académico universitario: una revisión sistemática", Más TIC, 2024 Publicación	<1 %
11	lisainsurtech.com Fuente de Internet	<1 %
12	Submitted to Universidad Carlos III de Madrid - EUR Trabajo del estudiante	<1 %
13	www.rediech.org Fuente de Internet	<1 %
14	www.spell.org.br Fuente de Internet	<1 %
15	simplified.com Fuente de Internet	<1 %
16	cuauhtemocgdl.edu.mx Fuente de Internet	<1 %
17	epsir.net Fuente de Internet	<1 %
18	Submitted to Universidad Privada Boliviana Trabajo del estudiante	<1 %

Excluir citas	Apagado	Excluir coincidencias	< 40 words
Excluir bibliografía	Activo		

Asesor

Dr. Kevin Rafael, Palomino Pacheco.

Agradecimientos

A nuestras familias, por ser nuestro mayor apoyo a lo largo de este camino. Por su amor, paciencia y confianza en nosotros que han sido nuestra mayor motivación para seguir adelante, gracias por cada palabra de aliento, por cada sacrificio y por estar siempre a nuestro lado, impulsándonos a alcanzar nuestras metas. Este logro es también suyo.

Índice

Índice de tabla

Índice de figura

Resumen

El presente estudio tiene como objetivo identificar señales de violencia verbal en comentarios y publicaciones en redes sociales mediante herramientas de ciencia de datos. Para ello, se utilizó un enfoque mixto (cuantitativo-cualitativo) con un diseño secuencial exploratorio, de tipo aplicado y alcance descriptivo. Se recopilaron y analizaron 2343 comentarios extraídos de *YouTube,* los cuales fueron procesados utilizando técnicas de minería de texto y aprendizaje automático. La clasificación de los comentarios en violentos y no violentos se llevó a cabo mediante el modelo *DistilBERT*, el cual alcanzó una precisión del 94.12%, una exactitud del 94.05% y un área bajo la curva (AUC) de 0.98.

Además, se diseñó una herramienta digital basada en *Django* y *Plotly* que permite visualizar patrones de violencia verbal en redes sociales mediante gráficos interactivos y reportes automatizados. La plataforma facilita la detección de términos agresivos y su distribución en los comentarios analizados, proporcionando un sistema eficiente para el monitoreo y análisis del lenguaje en entornos digitales. Los resultados evidenciaron que la violencia verbal en redes sociales se presenta principalmente en comentarios breves, con una alta concentración de términos agresivos. Asimismo, se identificaron limitaciones en la interpretación del contexto y en la clasificación de expresiones ambiguas, lo que sugiere la necesidad de mejorar la comprensión semántica del modelo.

El estudio contribuye al campo del PLN y la detección automatizada de discursos agresivos en redes sociales, proporcionando un marco metodológico sólido para el análisis de violencia verbal en entornos digitales.

Palabras clave: Análisis de sentimientos, violencia verbal, redes sociales, aprendizaje automático, procesamiento de lenguaje natural.

Abstract

This study aims to identify signs of verbal violence in comments and posts on social media using data science tools. A mixed-methods approach (quantitative-qualitative) was employed, following a sequential exploratory design with an applied, descriptive scope. A total of 2,343 comments extracted from YouTube were collected and analyzed using text mining and machine learning techniques. The classification of comments into violent and non-violent categories was performed using the DistilBERT model, which achieved a precision of 94.12%, an accuracy of 94.05%, and an area under the curve (AUC) of 0.98.

Additionally, a digital tool based on Django and Plotly was designed to visualize patterns of verbal violence on social media through interactive graphs and automated reports. This platform facilitates the detection of aggressive terms and their distribution in the analyzed comments, providing an efficient system for monitoring and analyzing language in digital environments.

The results showed that verbal violence on social media primarily appears in short comments with a high concentration of aggressive terms. Furthermore, limitations were identified in context interpretation and the classification of ambiguous expressions, suggesting the need to improve the model's semantic understanding. This study contributes to the field of natural language processing and the automated detection of aggressive discourse on social media, providing a solid methodological framework for analyzing verbal violence in digital environments.

Keywords: Sentiment analysis, verbal violence, social media, machine learning, natural language processing.

Introducción

En la era digital, el uso de redes sociales ha transformado la comunicación entre individuos, permitiendo una mayor interconectividad global. Sin embargo, esta misma apertura ha dado lugar a la proliferación de contenidos agresivos, entre ellos la violencia verbal. El crecimiento exponencial de plataformas como *Facebook*, *YouTube, Twitter* e *Instagram* ha facilitado la difusión de discursos dañinos, afectando la convivencia digital y generando consecuencias psicológicas adversas en los usuarios (Patchin & Hinduja, 2020). La Organización Mundial de la Salud (OMS, 2021) y organismos de derechos digitales han alertado sobre el impacto de este tipo de violencia en la salud mental, particularmente en jóvenes y adolescentes, poblaciones altamente expuestas a interacciones en línea.

A nivel nacional, el problema no es ajeno a la realidad social. En diversos estudios sobre comportamiento en redes sociales, se ha identificado un aumento significativo en el uso de insultos, discursos de odio y agresiones verbales en plataformas digitales (UNESCO, 2022). Esto no solo afecta a individuos específicos, sino que también deteriora la calidad del debate público y genera ambientes hostiles en el entorno digital (Citron, 2014). La falta de moderación eficaz y la dificultad para diferenciar entre comentarios inofensivos y expresiones de violencia verbal han hecho evidente la necesidad de desarrollar herramientas automatizadas para su detección y prevención (Waseem et al., 2017).

En este contexto, la presente investigación tiene como objetivo desarrollar un modelo basado en PLN y aprendizaje automático para la identificación y clasificación de señales de violencia verbal en redes sociales. Estudios previos han demostrado que las técnicas de PLN pueden ser herramientas efectivas para la detección de lenguaje agresivo en entornos digitales, con aplicaciones en la moderación de contenido y la prevención del ciberacoso (Schmidt & Wiegand, 2017). Este estudio es relevante no solo desde una perspectiva técnica, sino también desde un enfoque social, ya que la implementación de sistemas automatizados de moderación de contenido podría contribuir a la reducción del ciberacoso y a la mejora del bienestar digital.

La estructura de esta investigación se organiza en cinco capítulos. En el Capítulo 1, se presenta el planteamiento del estudio, donde se define el problema de investigación, su relevancia y los objetivos que guían el desarrollo del trabajo. Asimismo, se justifica la importancia del estudio y se establecen los alcances y limitaciones.

En el Capítulo 2, se expone el marco teórico, abordando las principales teorías y modelos conceptuales que sustentan la investigación. Se analizan enfoques provenientes del procesamiento de lenguaje natural, aprendizaje automático y análisis de discurso en redes sociales, proporcionando la base teórica para la detección de violencia verbal en entornos digitales.

El Capítulo 3 presenta las hipótesis y variables de estudio, definiendo con precisión la relación entre los elementos clave de la investigación. Se establecen las variables independientes y dependientes, así como su operativización para su medición y análisis.

En el Capítulo 4, se detalla la metodología del estudio, describiendo el enfoque mixto utilizado, el diseño de investigación, la recolección y procesamiento de datos, así como las técnicas de análisis aplicadas. También se explica la implementación del modelo *DistilBERT* y el desarrollo de la herramienta digital para la visualización de patrones de violencia verbal.

Finalmente, en el Capítulo 5, se presentan los resultados obtenidos, junto con las conclusiones derivadas del análisis. Además, se plantean recomendaciones para futuras investigaciones y aplicaciones prácticas del modelo propuesto, considerando las limitaciones identificadas y posibles mejoras en la detección automatizada de discursos agresivos en redes sociales.

Capítulo I: Planteamiento del estudio

1.1. Planteamiento y formulación del problema

1.1.1. Planteamiento del problema

Descripción del contexto

En la última década, el vertiginoso desarrollo de internet y de las redes sociales han creado nuevos e insospechados modos de circulación de la información y de estilos de sociabilidad (Lacunza et al., 2019), ello debido al creciente acceso a los dispositivos móviles, por ejemplo, según la Unión Internacional de Telecomunicaciones (UIT) en el año 2023, el 67% de la población mundial de diez años en adelante tenía un teléfono celular con acceso a internet (ONU, 2023), los cuáles no solo se usan para hacer llamadas sino con continuas visitas a las plataformas como *YouTube, Facebook, X, Instagram*, entre otros.

Esta adopción de la tecnología tan abrumadora y con una fascinación tan alta detrás de los medios digitales se veía que algunas cosas empezarían a cambiar, se presumía que el uso de tanta tecnología tendría aristas complicadas que las generaciones previas no tuvieron que afrontar (Kont, 2016). En tal sentido, hoy vivimos en un nuevo paradigma de vida y es considerar lo digital como parte de nuestras vidas. Es decir, las redes sociales son espacios y herramientas donde se produce todo tipo de actividades, muchas positivas. Sin embargo, muchas otras son negativas. La violencia es una de tantas y por ser producida en plataformas diseñadas para generar notoriedad, sus alcances son muchas veces mayores que la que se da en medios tradicionales.

Galván (2024) señala que,

> “Estos alcances se pueden dar a través del uso de palabras altisonantes, ofensivas y agresivas que puede tener un impacto significativo en la salud mental de quienes participan en estas interacciones”

Por ejemplo, un estudio realizado por Suler (2004), titulado *"The Online Disinhibition Effect"*, exploró; cómo el anonimato y la distancia física en línea

pueden llevar a un comportamiento más agresivo y desinhibido. Suler encontró que las personas tienden a sentirse menos restringidas en línea y, por lo tanto, más propensas a expresar emociones negativas y adoptar comportamientos agresivos.

Otro estudio realizado por Pieschl, Porsch y Kappich en 2018, titulado *"The relationship between adolescents' moral competencies and online aggression: Evidence from Germany"*, examinó la relación entre la competencia moral de los adolescentes y su participación en comportamientos agresivos en línea. Este estudio encontró que los adolescentes con niveles más bajos en una escala sobre moral eran más propensos a participar en comportamientos agresivos en línea (Galván, 2024).

Por otro lado, el uso de medios digitales y redes sociales va en aumento, generando un gran volumen de información accesible para el análisis, especialmente en el ámbito del monitoreo de medios. Este interés es relevante tanto para empresas como para gobiernos, que buscan entender las opiniones en las redes. Una técnica clave en este análisis es el "análisis de sentimiento o minería de opiniones", que consiste en el proceso de analizar y comprender las opiniones, actitudes y emociones de las personas hacia un tema o producto en particular; es decir, extraer la polaridad emocional de los textos. Esta técnica implica el uso de procesamiento de lenguaje natural, aprendizaje automático e inteligencia artificial para analizar los datos de texto de plataformas de redes sociales, *blogs* y foros en línea.

El objetivo principal del análisis de sentimientos es identificar la polaridad del texto, ya sea positiva, negativa o neutra (IAU, 2024). Por tal motivo, esta información puede proporcionar a los entes encargados de la salud mental valiosas ideas para tomar mejores decisiones en el diseño e implementación de políticas de prevención de la violencia (UNESCO,2018).

Peña (2021) señala que los métodos tradicionales para moderar contenido violento en redes sociales presentan importantes limitaciones. Por ejemplo, Frances Haugen, ingeniera de datos y exgerente de productos en *Facebook*, denunció que la plataforma prioriza contenido extremo y adictivo con el fin de mantener a los usuarios conectados durante el mayor tiempo posible. A pesar de

la existencia de mecanismos de moderación, los algoritmos de *Facebook* están diseñados para amplificar contenidos que generan controversia y emociones intensas, lo que incluye material violento y nocivo.

A pesar de haber contratado a más de cuarenta mil empleados para la moderación de contenido, este enfoque manual resulta insuficiente ante el abrumador volumen de publicaciones. Este proceso es lento y no escalable, lo que permite que el contenido violento se difunda antes de ser eliminado. Además, los algoritmos de aprendizaje automático implementados para detectar contenido inapropiado han demostrado ser ineficaces en la contención de la desinformación, como se evidenció durante las campañas antivacunación del COVID-19 en Estados Unidos. Estos sistemas enfrentan dificultades para comprender el contexto y los matices de los contenidos dañinos, lo que limita considerablemente su efectividad (Peña, 2021, pp. 269-270).

En consecuencia, la gestión manual de grandes volúmenes de información variable y dinámica para la clasificación sentimental se vuelve utópica. En este contexto, el PLN ha cobrado gran relevancia en los últimos años, constituyéndose como una de las áreas más significativas de la Inteligencia Artificial [IA] (Deng y Liu, 2018; McNamara et al., 2017). El PLN se centra en facilitar la comunicación entre personas y computadoras, permitiendo que estas últimas "entiendan" oraciones o textos (Cortez, Vega y Pariona, 2011). Dentro del PLN, se pueden desarrollar diversas aplicaciones, como el reconocimiento de voz, el entendimiento del habla, sistemas de diálogo y análisis sintáctico, entre otros (Deng y Liu, 2018). El análisis de sentimiento se erige como una subárea crucial de investigación dentro del PLN, consolidándose en la última década como un aspecto especial que toca múltiples dimensiones del procesamiento del lenguaje, incluyendo el análisis léxico, la resolución de correferencias, la desambiguación, el análisis de discurso, la extracción de información y el análisis semántico.

Dada la complejidad y la variabilidad del lenguaje utilizado en las redes sociales, se hace imprescindible el desarrollo de métodos más avanzados y eficientes para la detección de violencia verbal. No solo es fundamental para

proteger a los usuarios y prevenir situaciones de acoso, sino que también es vital para mejorar la seguridad en las plataformas digitales. En este sentido, las redes neuronales representan una herramienta moderna con un enorme potencial, ya que, mediante el aprendizaje profundo, pueden capturar patrones complejos en el lenguaje que escapan a los métodos tradicionales. La implementación de estas tecnologías avanzadas podría transformar significativamente la eficacia de los sistemas de moderación, permitiendo una respuesta más rápida y precisa ante contenido violento y perjudicial.

1.1.2. Formulación del problema

1.1.2.1 Problema General

¿Cómo se pueden detectar señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos?

1.1.2.2 Problemas específicos

- ¿Qué características y patrones de lenguaje están presentes en comentarios y publicaciones que contienen violencia verbal en redes sociales?
- ¿Cómo se puede desarrollar un modelo de análisis de sentimientos basado en técnicas de procesamiento de lenguaje natural (PLN) para detectar señales de violencia verbal en redes sociales?
- ¿Qué nivel de precisión, sensibilidad y especificidad puede alcanzar el modelo de análisis de sentimientos para detectar violencia verbal?
- ¿Qué tipo de informes y visualizaciones podrían ser útiles para identificar tendencias y patrones de violencia verbal en redes sociales, y cómo pueden implementarse alertas automáticas para su monitorización?

1.2. Determinación de objetivos

1.2.1. Objetivo General

Identificar señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos.

1.2.2. Objetivos Específicos

- Identificar y analizar las características y patrones de lenguaje en comentarios de redes sociales que contienen violencia verbal, mediante la extracción, etiquetado y preprocesamiento de un conjunto de datos.

- Desarrollar un modelo de análisis de sentimientos para detectar violencia verbal con técnicas de PLN.

- Evaluar y optimizar el rendimiento del modelo con métricas de clasificación.

- Diseñar una herramienta digital que genere informes y ventanas emergentes que muestren patrones de violencia verbal.

1.3. Justificación e importancia del estudio

1.3.1. Justificación teórica

El análisis de sentimientos en redes sociales se ha vuelto crucial para entender las interacciones y emociones que se desarrollan en el entorno digital. La violencia verbal, que se manifiesta a través de insultos y acosos en línea, es un problema serio que afecta el bienestar de muchas personas. En tal sentido, el presente estudio es importante por varias razones:

- Importancia Social. A medida que la comunicación en línea se intensifica, es fundamental identificar y comprender las señales de violencia verbal. El análisis de sentimientos permite detectar estos patrones, ayudando a crear espacios digitales más seguros y saludables.

- Innovaciones Tecnológicas. Las redes neuronales han transformado el campo del procesamiento del lenguaje natural, mejorando la manera en que analizamos el lenguaje. Estos modelos son capaces de captar las sutilezas del lenguaje humano, lo que permite un

análisis más profundo y preciso de las emociones expresadas en las redes sociales.

- Aporte al Conocimiento. Este estudio no solo busca ampliar el entendimiento sobre el análisis de sentimientos, sino también su aplicación en la detección de la violencia verbal. Al combinar enfoques avanzados en procesamiento de lenguaje y redes neuronales, se pretende desarrollar un modelo que sea más efectivo y útil para investigadores y profesionales en el ámbito de la seguridad en línea.

- Relevancia Práctica. Los hallazgos de esta investigación tendrán implicaciones no solo en el ámbito académico, sino también en la creación de herramientas que ayuden a moderar y prevenir el acoso en las redes sociales, contribuyendo así a políticas más efectivas para combatir la violencia en línea.

1.3.2. Justificación metodológica

La metodología elegida para este estudio es clave para abordar de manera efectiva el análisis de sentimientos y la detección de la violencia verbal en redes sociales. A continuación, se explican las razones detrás de esta elección:

- Técnicas de Análisis Efectivas. Optamos por utilizar redes neuronales, especialmente los modelos de aprendizaje profundo, porque son ideales para manejar grandes cantidades de datos y reconocer patrones complejos en el lenguaje. Estas herramientas han demostrado ser muy eficaces en el procesamiento del lenguaje natural, lo que nos permitirá realizar un análisis más preciso de las emociones y detectar señales de violencia verbal.

- Limpieza y Preparación de Datos. La metodología incluye un exhaustivo proceso de preprocesamiento de datos, fundamental para limpiar y organizar la información extraída de las redes sociales. Este paso es crucial para eliminar elementos irrelevantes y asegurar que

los modelos se entrenen con datos de alta calidad, lo que mejora la fiabilidad de nuestros resultados.

- Evaluación Rigurosa. Implementaremos técnicas de validación cruzada y utilizaremos métricas como la precisión, el *recall* y la F1-*score* para evaluar la efectividad del modelo. Esto garantiza que los resultados sean sólidos y aplicables a otros contextos, lo cual es esencial en cualquier investigación que aspire a hacer contribuciones significativas al campo.

- Comprensión del Contexto. Nuestra metodología también contempla un análisis cualitativo de los resultados. Esto nos permitirá no solo presentar datos numéricos, sino también interpretar el contexto detrás de ellos, lo cual es vital para entender la violencia verbal en el entorno digital y las dinámicas sociales que la rodean.

- Relevancia Práctica. Al adoptar un enfoque metodológico riguroso y actualizado, buscamos no solo contribuir al conocimiento académico, sino también ofrecer herramientas prácticas que puedan ser útiles para plataformas de redes sociales y organizaciones que trabajan en la prevención de la violencia verbal.

1.3.3. Justificación social

El análisis de sentimientos y la detección de la violencia verbal en redes sociales son fundamentales por diversas razones sociales que destacan su importancia:

- Salud Mental de las Personas. La violencia verbal en línea puede afectar gravemente la salud mental de quienes la sufren. Este estudio busca contribuir a la creación de un entorno digital más saludable, protegiendo el bienestar emocional de los usuarios y ayudando a reducir el daño que provoca el acoso en línea.

- Fomento de la Convivencia Pacífica. Al identificar y abordar conductas abusivas de manera temprana, se pueden evitar conflictos

mayores y fomentar una convivencia más armónica en las plataformas digitales. Esto es especialmente relevante en una era donde la comunicación en línea es tan común.

- Conciencia sobre el Problema. Al investigar y dar visibilidad a la violencia verbal en redes sociales, se genera mayor conciencia sobre esta problemática. Esto puede motivar a comunidades, plataformas y autoridades a implementar medidas más efectivas para combatir el acoso y promover un uso más responsable de las redes.

- Empoderamiento de Usuarios. La investigación puede proporcionar a los usuarios herramientas para reconocer y reportar comportamientos abusivos. Esto no solo les da poder, sino que también fomenta un ambiente donde se alienta a denunciar y a proteger los derechos de todos en la comunidad digital.

- Influencia en Políticas Públicas. Los resultados de este estudio pueden servir como base para que instituciones y responsables de políticas comprendan la gravedad de la violencia verbal en línea, ayudando en la creación de normativas y programas que fomenten la seguridad y el respeto en las interacciones digitales.

1.4. Limitaciones de la presente investigación

- Datos limitados en español. Existe menor disponibilidad de conjuntos de datos de calidad en español, lo que complica la tarea de entrenar modelos robustos para detectar violencia verbal en redes sociales. Los datasets que contienen ejemplos de interacciones violentas en español pueden ser más difíciles de conseguir, pueden estar mal estructurados, mal etiquetados o no existir.

- Diferencias lingüísticas. La mayoría de los modelos de procesamiento de lenguaje natural están entrenados principalmente en inglés. Esto puede afectar la precisión de los análisis en español, ya que muchas palabras, expresiones y modismos no tienen traducciones directas o

cambian de significado según el contexto. Esto también incluye las diferencias regionales dentro del mismo idioma, lo que dificulta la identificación precisa de la violencia verbal en diferentes dialectos del español que se hablan en Perú.

- Limitaciones de memoria y almacenamiento. El trabajo con grandes conjuntos de datos textuales y modelos complejos puede requerir mucha memoria RAM y almacenamiento en disco. Lo que podría ocasionar que nos enfrentemos a cuellos de botella durante el procesamiento de datos o la ejecución de modelos.

- Costo de infraestructura en la nube. Otras limitaciones que se puede presentar son los costos elevados de infraestructura en la nube (como AWS, Google Cloud, o Azure), ya que estos servicios pueden ser costosos, especialmente para el entrenamiento continuo de modelos de redes neuronales o para procesar grandes volúmenes de datos en tiempo real.

- Sobreajuste *(overfitting).* Los modelos más complejos, como redes neuronales con muchas capas o con un gran número de parámetros, son más propensos al sobreajuste. Esto ocurre cuando el modelo se ajusta demasiado bien a los datos de entrenamiento, pero pierde capacidad de generalización en datos nuevos. Esto es especialmente crítico en el análisis de violencia verbal, donde el lenguaje puede variar significativamente en diferentes contextos y plataformas de redes sociales.

- Dificultad en la interpretabilidad del modelo. A medida que los modelos se vuelven más complejos, se vuelven más difíciles de interpretar. Esto es una limitación cuando necesitas explicar los resultados a un público no técnico o cuando es necesario entender cómo el modelo toma decisiones. Los modelos como las redes neuronales profundas son conocidos por ser cajas negras, lo que

podría ser un problema si necesitamos justificar por qué un comentario fue clasificado como violento.

Capítulo II: Marco teórico

2.1 Antecedentes de la investigación

2.1.1. Internacionales

A nivel internacional, destaca el estudio realizado en 2018 por Paul David Cumba Armijos el cual llevó a cabo una investigación titulada *"Predicción de ataques de cyber bullying mediante técnicas de aprendizaje profundo apoyándose en un corpus de entrenamiento para la clasificación de texto en español".* El objetivo principal fue desarrollar un modelo de predicción de ataques de cyberbullying utilizando aprendizaje profundo, apoyado en un corpus en español diseñado para clasificar e identificar textos con características de bullying. Para lograr este objetivo, se emplearon técnicas de aprendizaje profundo, específicamente una red neuronal convolucional (CNN). El proceso comenzó con la recolección y preparación de un corpus de texto en español, en el cual se etiquetaron los datos como "*bullying*" o "*no bullying*". Este corpus fue preprocesado mediante técnicas de limpieza, eliminación de ruido y normalización de datos. La muestra de estudio incluyó 800,000 usuarios activos de la red social X (anteriormente *Twitter*) en Ecuador, de donde se extrajeron un total de 83,400 *tweets* mediante palabras clave.

Los resultados mostraron que los modelos generados en diferentes iteraciones de validación presentaron baja varianza, lo que indica consistencia y estabilidad en el rendimiento del modelo. Además, el modelo alcanzó un alto porcentaje de precisión en cada iteración. En cuanto a las probabilidades de clasificación, se obtuvo un 98.19% de probabilidad de identificar un texto como "*bullying*" y un 99.66% como "no *bullying*". La precisión global del modelo fue del 98.85%, con un porcentaje de error de predicción de solo el 1.15%, lo cual evidencia su eficacia para identificar textos asociados al cyberbullying. Por otro lado, los resultados sugieren que el modelo tenía una mayor probabilidad de predecir un tweet como "no *bullying*" en comparación con la categoría "*bullying*". Esto se atribuye a la mayor cantidad de ejemplos etiquetados como "no *bullying*" en el corpus, lo que permitió al modelo aprender patrones de lenguaje sin signos de *bullying* con mayor frecuencia. No obstante, la diferencia en la tasa de aciertos

entre ambas categorías fue mínima, lo que indica un buen equilibrio en el desempeño del modelo para clasificar correctamente ambos tipos de textos.

Siguiendo esta misma línea, en 2019, Daniel José Silva Oliveira, Paulo Henrique de Souza Bermejo, José Roberto Pereira y Daniely Aparecida Barbosa realizaron un estudio titulado "La aplicación de la técnica de análisis de sentimiento en medios sociales como instrumento para las prácticas de la gestión social a nivel gubernamental". El objetivo de esta investigación fue analizar las opiniones de los ciudadanos expresadas en *Twitter* sobre algunos de los principales programas sociales en Brasil durante el gobierno de Dilma Rousseff. La metodología empleada fue de carácter aplicado, interdisciplinario y exploratorio, con un enfoque mixto: cualitativo en el proceso de clasificación de los datos y cuantitativo al verificar la frecuencia de opiniones positivas, negativas y neutrales en la base de datos. El proceso metodológico incluyó varios pasos: la definición del objeto de estudio, la selección de la fuente de datos, la definición de palabras clave para la minería de datos, la elección de una aplicación de minería de opiniones, la recopilación y preparación de datos, la delimitación del conjunto de entrenamiento, la clasificación automatizada y la validación de los resultados.

En cuanto a la muestra, tras aplicar filtros para eliminar *tweets* duplicados y reducir el ruido, se seleccionaron 10176 *tweets* de un total inicial de 59012. Los resultados del análisis de sentimiento revelaron que en los programas sociales "Pronatec" y "Minha Casa, Minha Vida" predominaban las opiniones positivas, mientras que en los programas "Mais Médicos" y "Bolsa Família" se observaba una mayoría de opiniones negativas. Este análisis de sentimiento proporciona una herramienta valiosa para la gestión social gubernamental, al facilitar la comprensión de las percepciones ciudadanas sobre programas sociales clave.

Por otro lado, en 2021, Nirmal Varghese Babu y E. Grace Mary Kanaga publicaron un estudio titulado "Análisis de sentimientos en datos de redes sociales para la detección de depresión usando inteligencia artificial: Una revisión". El objetivo de esta investigación fue revisar el uso de análisis de sentimientos en redes sociales para la detección de depresión mediante diversas técnicas de inteligencia artificial.

La metodología se basó en una exhaustiva revisión de 101 artículos científicos de revistas especializadas como IEEE, *Springer* y ACM, que abordan temas de análisis de datos, redes sociales, procesamiento de lenguaje natural, análisis de sentimientos y detección de depresión. Los aspectos clave examinados en los estudios revisados incluyen:

- Análisis de Sentimientos. Se analizaron diferentes métodos para determinar la polaridad (positiva, negativa o neutral) y para identificar emociones específicas como alegría, tristeza y enojo en textos de redes sociales.

- Extracción de Características. Se utilizaron técnicas para convertir texto en datos numéricos, empleando métodos como TF-IDF, bolsas de palabras, N-gramas y herramientas de PLN como Word2Vec, GloVe y SentiWordNet.

- Clasificación Multiclase. Se revisaron enfoques de clasificación detallada, que no solo dividían los textos en positivos y negativos, sino que los organizaban en varias categorías emocionales. Para esto, se emplearon algoritmos como máquinas de soporte vectorial, árboles de decisión, Random Forest, redes neuronales convolucionales (CNN) y redes LSTM.

- Fuentes de Datos: Se recopilaron textos, emoticonos y emojis de diversas plataformas de redes sociales, incluyendo *Twitter, Facebook*, *Reddit* y *Weibo*, que se analizaron para obtener una visión completa del estado emocional de los usuarios.

Los resultados de esta revisión destacaron que el uso de modelos de clasificación detallados, que identifican múltiples emociones en lugar de limitarse a polaridades simples, mejoró significativamente la precisión en la detección de sentimientos asociados a la depresión. En particular, las redes neuronales profundas, como las CNN y las LSTM, demostraron ser especialmente efectivas en esta tarea. Además, se resaltó la importancia de considerar los emoticonos y emojis

como complementos del análisis textual, ya que ofrecen valiosas pistas sobre las emociones subyacentes de los usuarios.

La revisión concluyó que los métodos de aprendizaje profundo, especialmente la combinación de CNN y LSTM, superan a los enfoques tradicionales en la clasificación de sentimientos en el contexto de la detección de depresión. También se confirmó que redes sociales como *Twitter* y *Facebook* son fuentes útiles para identificar señales tempranas de depresión, permitiendo un monitoreo emocional que puede ayudar a detectar casos que requieran atención profesional.

Asimismo, en un estudio titulado "Detección de expresiones de incitación a la violencia en tweets en urdu 2022 utilizando redes neuronales convolucionales", los autores Muhammad Shahid Khan, Muhammad Shahid Iqbal Malik y Aamer Nadeem presentan un marco de inteligencia artificial dirigido a identificar expresiones violentas en publicaciones en urdu. El objetivo de este trabajo fue desarrollar un sistema robusto para detectar incitación a la violencia en tweets mediante redes neuronales convolucionales (CNN). El estudio recopiló un corpus de 4808 tweets obtenidos manualmente de cuentas de *Twitter* en Pakistán entre febrero de 2018 y junio de 2022. Los tweets se clasificaron en dos categorías: aquellos que incitaban a la violencia (2402 tweets) y aquellos que no lo hacían (2402 tweets). Utilizando modelos de aprendizaje profundo, entre ellos una CNN unidimensional (1D-CNN), los investigadores compararon su rendimiento con modelos como urdu-BERT, Urdu-RoBERTa, BiLSTM y otros métodos de aprendizaje automático. La efectividad de estos modelos se evaluó mediante métricas de precisión, macro F1-score y AUC.

Los resultados destacaron la eficacia de la arquitectura 1D-CNN, la cual alcanzó una precisión del 89.84% en la detección de incitación a la violencia, superando a las redes neuronales recurrentes y a los algoritmos tradicionales de aprendizaje automático. Este rendimiento se atribuye a la capacidad de la 1D-CNN para capturar patrones tanto locales como globales en los datos, lo que mejora su generalización y la convierte en una herramienta potente para moderar contenido

en línea. Además, el modelo demostró ser capaz de identificar patrones de incitación que otros modelos no lograban captar.

Si bien los modelos preentrenados, como urdu-BERT y Urdu-RoBERTa, fueron útiles para comprender el contexto semántico, no superaron a la 1D-CNN en la tarea específica de detectar incitación a la violencia, lo cual subraya la idoneidad de la 1D-CNN para este tipo de análisis. Este estudio contribuye así al campo de la moderación de contenido en redes sociales, proporcionando un enfoque especializado y efectivo para la detección de expresiones de incitación en entornos digitales.

Por otro lado, el artículo titulado "Machine learning para la detección de situaciones de riesgo a través del uso de aplicaciones de citas entre estudiantes universitarios, 2023" fue elaborado por Mariana Carolyn Cruz Mendoza, Roberto Ángel Meléndez Armenta y Narendra Velázquez Carmona. Su objetivo fue analizar la información compartida en aplicaciones de citas, como Tinder, para identificar patrones de comportamiento y características de personalidad violenta en los perfiles de usuario, con el fin de prevenir situaciones de riesgo y violencia entre jóvenes universitarios. La metodología empleada incluye un enfoque cuantitativo basado en encuestas dirigidas a estudiantes del Tecnológico de Estudios Superiores de Valle de Bravo, abordando aspectos como el uso, percepción de seguridad y experiencias de riesgo en aplicaciones de citas. Los datos se procesan mediante herramientas de software libre, como Python y Anaconda, y se someten a técnicas de limpieza, exploración y visualización. Además, se aplican técnicas de machine learning para analizar patrones de comportamiento violento. El modelo se entrena usando datos históricos para clasificar y predecir perfiles de riesgo, lo que permitirá, en futuras fases del proyecto, el desarrollo de una red neuronal que clasifique los perfiles de usuario en función de características asociadas a comportamientos violentos. Se consideran variables como el tipo de discurso de odio o la dependencia emocional, con el objetivo de identificar patrones específicos de agresores y víctimas.

La muestra utilizada está compuesta por estudiantes de diez programas académicos de nivel universitario. Los resultados parciales indican que la mayoría

de los usuarios usa las aplicaciones de citas de forma ocasional y, aunque están informados sobre las medidas de seguridad, pocos actúan ante situaciones de violencia en línea, resaltando la necesidad de proporcionar orientación y apoyo práctico. Asimismo, se identificaron clasificaciones de comportamiento violento que se utilizarán en la red neuronal para mejorar la detección de perfiles de riesgo y así contribuir a la prevención de violencia en el contexto universitario.

Los antecedentes internacionales presentados anteriormente se relacionan estrechamente con la investigación en curso sobre la identificación precisa y eficiente de señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos. En particular, estos estudios aportan enfoques metodológicos como la extracción y preprocesamiento de datos, el desarrollo de modelos de análisis de sentimientos basados en técnicas de PLN, y la evaluación de estos modelos con métricas de clasificación. Además, resaltan la importancia de diseñar herramientas prácticas, como aplicaciones que generen informes y patrones, contribuyendo directamente a los objetivos específicos de esta investigación.

2.1.2. Nacionales

Desde el punto de vista nacional, se encuentra un estudio propuesto por Gabriel Hélard Zárate Calderón en 2019 en el cual se desarrolló una investigación cuyo objetivo fue crear un modelo de análisis de sentimientos en medios periodísticos y redes sociales, implementando redes neuronales recurrentes y evaluando las capacidades de modelos basados en transformers. La investigación se centró en comparar diversas técnicas de procesamiento y analizar el rendimiento de los modelos en la detección de sentimientos en español. La metodología empleada incluyó el entrenamiento de múltiples modelos de redes neuronales recurrentes y transformers, utilizando optimizadores y planificadores específicos *(schedulers)* para mejorar la precisión en el análisis de sentimientos. Los datos fueron extraídos de redes sociales y medios periodísticos, destacándose el dataset TASS 2019 y un conjunto de datos proporcionado por una empresa de monitoreo de medios, ambos en idioma español.

Aunque el tamaño de la muestra no fue detallado, se especifica que incluye tanto textos de redes sociales como artículos de medios de comunicación. Los resultados obtenidos muestran que el mejor modelo alcanzó una precisión del 88.39% para el análisis de sentimientos en artículos periodísticos y del 76.42% en *tweets*. Estos resultados subrayan la eficacia del modelo para interpretar y clasificar sentimientos en diferentes tipos de textos, siendo especialmente eficiente en el análisis de contenido periodístico.

Del mismo modo, Gustavo Adolfo Reyes-Paredes tuvo como objetivo demostrar que la exposición a noticias positivas mejora el estado de ánimo de las personas. Para ello, se empleó una metodología basada en un modelo de red neuronal LSTM para clasificar noticias en positivas o negativas y, posteriormente, se realizó un experimento utilizando el test PANAS para evaluar el estado de ánimo de los participantes antes y después de leer noticias positivas. La población de estudio incluyó noticias de medios peruanos y 520 personas en el experimento, con una muestra de 20,000 noticias (10,000 positivas y 10,000 negativas) y 520 participantes entre 20 y 45 años. Los resultados mostraron que el modelo LSTM logró una precisión del 87.98% en la clasificación de noticias, y el test PANAS indicó que la lectura de noticias positivas efectivamente mejoró el estado de ánimo de los participantes.

En 2021, Josue Angel Mauricio Salazar presenta un estudio el cual tiene como objetivo desarrollar un modelo de análisis de sentimientos para identificar discurso de odio en tweets, empleando algoritmos de procesamiento de lenguaje natural, específicamente Word2Vec y un modelo de redes neuronales recurrentes con LSTM bidireccional. La metodología se basó en el uso de PLN y técnicas de aprendizaje profundo, generando representaciones vectoriales de palabras mediante Word2Vec y entrenando una red neuronal recurrente bidireccional con LSTM para clasificar los tweets según su contenido de odio. La población del estudio consistió en un conjunto de datos de 6,000 tweets etiquetados, aunque no se especifica una muestra, por lo que se asume que se utilizó la totalidad de la población para el análisis. Los resultados indican que el modelo fue capaz de clasificar los tweets de manera efectiva, utilizando métricas de evaluación que evidencian su capacidad para identificar discursos de odio en redes sociales.

Para el 2022, un estudio presentado por Franklin Barrientos Porras, tuvo como objetivo identificar la correlación entre la opinión pública en *Twitter*, medida mediante análisis de sentimientos, y la movilidad poblacional durante el periodo de cuarentena en Perú. La metodología empleada fue un diseño ecológico basado en el análisis de datos secundarios de acceso público, centrado en publicaciones realizadas en Twitter durante el tiempo de cuarentena. La población consistió en todos los tweets que cumplieran con los criterios de inclusión y exclusión establecidos para el periodo de cuarentena en Perú, de los cuales, luego de aplicar estos criterios, se analizaron 89,928 tweets. Los resultados mostraron una percepción pública principalmente negativa hacia la cuarentena, con un 84% de los tweets clasificados como negativos. También se observó una relación entre la opinión pública expresada en *Twitter* y los patrones de movilización en el país.

Finalmente, Marco Antonio Adrianzen Zapata y Juan Fernández Olaya en 2023 realizaron una investigación cuyo objetivo fue implementar un modelo de inteligencia artificial para analizar sentimientos y emociones en las redes sociales de los equipos de la Liga 1 de fútbol peruano. Dentro de sus objetivos específicos, el estudio buscó desarrollar un modelo de aprendizaje automático que analizara y clasificara los sentimientos en publicaciones y comentarios, recopilar un conjunto de datos representativo de emociones, identificar patrones de expresión de sentimientos y mejorar el rendimiento del modelo. La metodología fue de tipo aplicada y empleó un diseño experimental. Las técnicas incluyeron web scraping para recolectar datos de redes sociales y el uso del algoritmo Naïve Bayes para el análisis de sentimientos. La población se definió como los 18 equipos participantes en la Liga 1 de Perú en el 2023, y la muestra se centró en los equipos más populares, seleccionados en función de la cantidad de seguidores e interacción en sus publicaciones durante el segundo semestre del año. Los resultados revelaron que, en general, los sentimientos expresados en redes sociales sobre estos equipos fueron mayormente positivos, con un 33.47% de comentarios clasificados como positivos, un 38.33% como neutros y un 28.20% como negativos. Este análisis sugiere una tendencia positiva en la interacción de los fanáticos hacia sus equipos en redes sociales, reflejando una percepción favorable en gran parte de los comentarios evaluados.

Los antecedentes nacionales presentados anteriormente se relacionan estrechamente con la investigación en curso sobre la identificación de señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos. Estos estudios aportan enfoques metodológicos relevantes, como el uso de técnicas avanzadas de PLN, modelos de aprendizaje profundo (LSTM, transformers) y herramientas de análisis de sentimientos. Además, destacan el empleo de datasets específicos y estrategias de evaluación para optimizar el rendimiento de los modelos, lo que contribuye directamente a los objetivos del presente estudio: desde la extracción y preprocesamiento de datos, hasta el desarrollo, evaluación y diseño de herramientas digitales que permitan identificar patrones de violencia verbal de manera precisa y eficiente.

2.2. Bases teóricas

2.2.1 Desarrollo histórico

El análisis de sentimientos (figura 1) se originó en diversas áreas como la psicología, la lingüística y la informática. No obstante, fue con el avance de la era digital y la popularidad de las redes sociales cuando realmente se consolidó como un campo independiente de estudio. En los inicios de la informática, los investigadores se enfocaron en la idea de enseñar a las máquinas a comprender y manejar el lenguaje humano. Este ámbito de estudio se conoce como procesamiento del lenguaje natural (PNL), el cual inició en la idea de crear una máquina traductora durante la Segunda Guerra Mundial, en la década de 1940. La idea original era convertir un idioma a otro, pero usando el cerebro de las computadoras; sin embargo, después de eso surgió la idea de convertir el lenguaje humano en lenguaje informático y al revés (Lisa, 2022).

En la década de 1980, los investigadores comenzaron a desarrollar algoritmos para analizar el tono y la subjetividad en el texto. Estos primeros trabajos sentaron las bases para lo que eventualmente se convertiría en el análisis de

sentimientos (Sánchez, 2023). En los años 2000, con la aparición de las redes sociales es donde llega el auge del análisis de sentimientos gracias a plataformas como *Facebook* y *Twitter* que generaron enormes cantidades de datos textuales no estructurados. Esto proporcionó una base para aplicar análisis de sentimientos a gran escala.

Con el crecimiento exponencial de las redes sociales, se veía que algunas cosas empezarían a cambiar (International Republican Institute, 2016) y se evidenció la necesidad de detectar patrones de lenguaje abusivo, incluyendo la violencia verbal. Durante esta época, surgieron varios estudios sobre la detección automática de ciberacoso, odio y agresión en redes sociales, empleando técnicas de PLN. Estas investigaciones utilizaron enfoques basados en palabras clave, diccionarios predefinidos y análisis semántico básico para identificar lenguaje ofensivo.

Con el desarrollo de modelos avanzados de deep learning, como los modelos de redes neuronales recurrentes (RNN) y transformadores (como BERT y GPT), se dio un gran salto en el análisis de texto y la comprensión del lenguaje natural. Estos modelos permitieron una mayor precisión en la detección de emociones y actitudes, incluyendo la violencia verbal, al tener en cuenta el contexto y el tono del mensaje.

Actualmente, los esfuerzos de investigación se enfocan en mejorar la precisión y eficiencia de los modelos para detectar violencia verbal en tiempo real en redes sociales. Sin embargo, el desafío persiste en lograr una detección efectiva que sea también eficiente en términos de recursos computacionales. A medida que las redes sociales continúan evolucionando y se generan nuevos tipos de interacciones, la detección de violencia verbal se ha convertido en un área crítica de investigación para prevenir el ciberacoso y otras formas de agresión en línea.

Figura 1

Desarrollo histórico del análisis de sentimientos

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

2.2.2 Fundamentación teórica

La fundamentación teórica de esta investigación (Figura 2) aborda un marco multidisciplinario que integra teorías provenientes de diversas áreas del

conocimiento, tales como la lingüística computacional, el PLN, la ciencia de datos y el análisis de comportamiento en redes sociales. Este marco conceptual permite estructurar el estudio y justificar las metodologías utilizadas para la detección de violencia verbal en plataformas digitales.

A. Teorías sobre Violencia y Discurso en Redes Sociales

El presente estudio se fundamenta en teorías sociológicas y psicolingüísticas sobre la violencia verbal en entornos digitales. La Teoría del Aprendizaje Social (Bandura, 1977) sostiene que la conducta violenta se aprende a través de la observación e imitación de otros, lo que en el contexto de redes sociales se traduce en la reproducción de discursos agresivos dentro de comunidades en línea. Asimismo, la Teoría del Comportamiento Desinhibido en Línea (Suler, 2004) explica cómo la anonimidad y la ausencia de consecuencias en entornos digitales fomentan la expresión de discursos violentos.

Por otro lado, el Modelo de Análisis Crítico del Discurso (Fairclough, 1992) aporta un enfoque lingüístico para examinar la construcción del discurso violento en redes sociales, permitiendo identificar estructuras narrativas que promueven la agresión y el odio. Además, la Teoría de la Incivilidad en la Comunicación Digital (Papacharissi, 2004) permite contextualizar la agresividad en línea como una expresión de polarización ideológica y discursos de odio.

B. Modelos de PLN y Aprendizaje Automático

Dado que la investigación utiliza técnicas avanzadas de PLN para la detección de violencia verbal, es fundamental la inclusión de teorías relacionadas con el análisis automático del lenguaje. El Modelo de Representación Distribuida del Lenguaje (Mikolov et al., 2013) introdujo técnicas como Word2Vec, que permiten representar palabras en vectores numéricos de alta dimensión, capturando sus relaciones semánticas y contextuales. Asimismo, el Modelo de Transformadores (Vaswani et al., 2017), sobre el cual se basa BERT y DistilBERT, es esencial en este estudio, ya que permite el análisis bidireccional del contexto lingüístico, mejorando la precisión en la clasificación de texto.

La investigación también se sustenta en el Enfoque de Aprendizaje Profundo en NLP (Young et al., 2018), que establece cómo las redes neuronales profundas pueden capturar patrones complejos en el lenguaje natural, facilitando la clasificación de comentarios agresivos con alto nivel de precisión. Dentro de este marco, el Modelo DistilBERT (Sanh et al., 2019) se presenta como una solución optimizada y eficiente para la clasificación de texto con menos recursos computacionales.

C. Ciencia de Datos y Evaluación del Rendimiento del Modelo

El estudio se apoya en la Teoría del Aprendizaje Supervisado (Bishop, 2006), que explica cómo un modelo de machine learning aprende patrones a partir de datos etiquetados. También se incorpora la Metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) (Wirth & Hipp, 2000), un enfoque estructurado que guía el proceso de minería de datos desde la recolección hasta la evaluación del modelo.

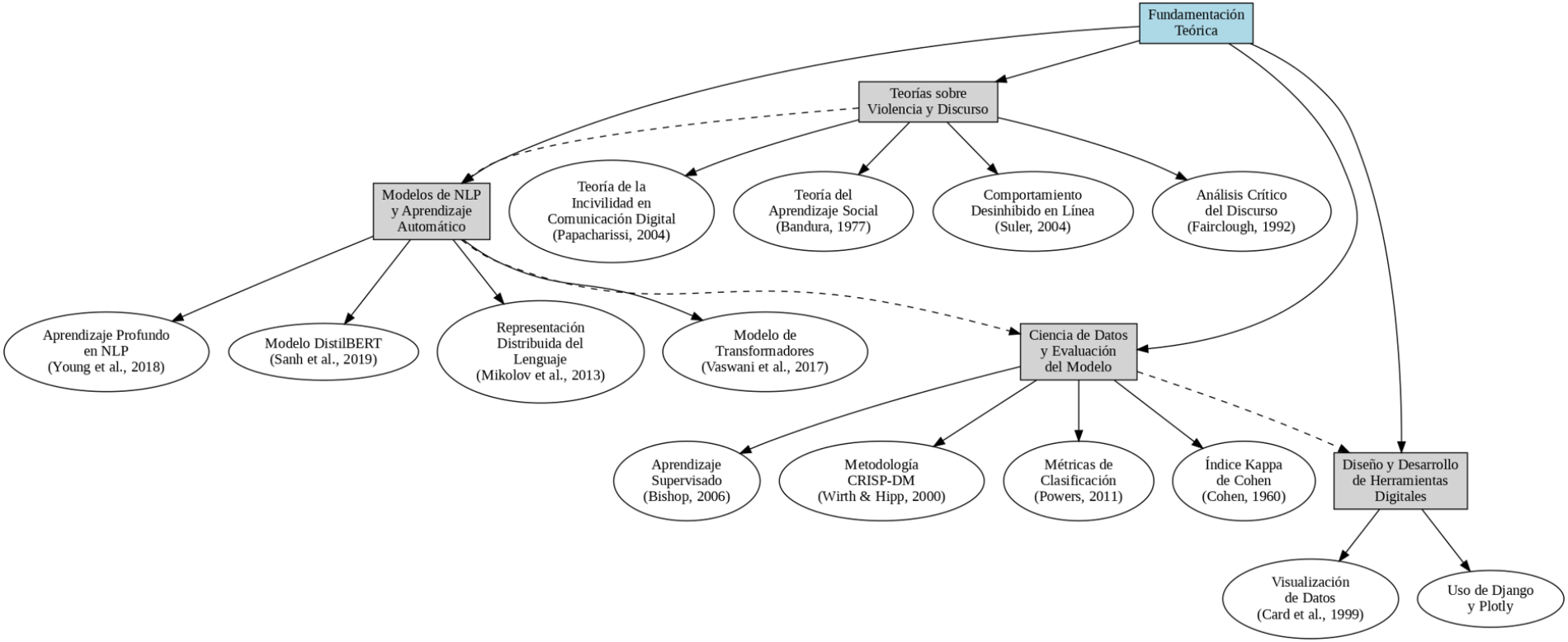
Para la evaluación del desempeño del modelo, se consideran métricas clásicas de clasificación como la precisión, exhaustividad y F1-Score, las cuales se derivan del análisis de la matriz de confusión (Powers, 2011). Además, el Índice Kappa de Cohen (Cohen, 1960) permite medir la concordancia entre las predicciones del modelo y las etiquetas reales, ajustado por el azar.

D. Diseño y Desarrollo de Herramientas Digitales para la Visualización de Datos

La última parte del presente estudio, que corresponde a la construcción de una herramienta digital para la detección de violencia verbal, se fundamenta en el paradigma de la Visualización de Datos (Card et al., 1999). Este enfoque sostiene que la representación gráfica de los datos facilita su interpretación y la toma de decisiones informadas. En este sentido, la utilización de frameworks como Django y Plotly proporcionan una interfaz interactiva y comprensible para el usuario.

Figura 2

Fundamentación teórica



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

2.2.3 Marco conceptual

Esta parte de la investigación proporciona una estructura organizada de los conceptos y variables clave que fundamentan la investigación. Se basa en la fundamentación teórica previamente establecida y tiene como objetivo guiar el análisis de los datos, asegurando la coherencia metodológica y conceptual a lo largo del estudio. El marco conceptual está conformado por dos variables principales: la primera, comentarios y publicaciones en redes sociales, corresponde a la variable independiente; mientras que la segunda, categoría del sentimiento de violencia verbal, representa la variable dependiente.

A. Definición conceptual de la variable 1: Comentarios y publicaciones de redes sociales.

Los comentarios en redes sociales son respuestas o reacciones generadas por los usuarios a publicaciones, imágenes o videos compartidos en plataformas como *Facebook, Instagram o Twitter. Sirven* como punto de partida para conversaciones digitales, permitiendo a las personas expresar sus pensamientos, opiniones o apreciaciones directamente debajo del contenido con el que interactúan. (Simplified., s.f.)

B. Definición conceptual de la variable 2: categoría del sentimiento de violencia verbal.

En el contexto del análisis de sentimientos aplicado a la violencia verbal en redes sociales, las categorías del sentimiento se refieren a las distintas clasificaciones que se utilizan para identificar y evaluar las emociones expresadas en los comentarios o publicaciones. Estas categorías permiten determinar si un mensaje contiene elementos de agresión, hostilidad o abuso verbal. Por ejemplo, se pueden identificar diversos tipos como burlas, uso de ofensas, uso de sobrenombres, entre otros. En el presente estudio las categorías son si contiene violencia (1) y no contiene violencia (0) (Oficina para la Salud de la Mujer, 2023).

2.3 Definición de términos básicos

2.3.1 Violencia Verbal

La violencia verbal se refiere al uso de lenguaje ofensivo, abusivo o denigrante con el fin de herir, intimidar o manipular a otra persona. En el contexto de las redes sociales, se manifiesta en comentarios, mensajes o publicaciones que contienen insultos, amenazas o cualquier expresión diseñada para causar daño emocional o psicológico.

2.3.2 Redes Sociales

Las redes sociales son plataformas en línea que permiten la creación, intercambio y difusión de contenido generado por los usuarios. Entre las más populares se encuentran *Facebook, Twitter, Instagram* y *TikTok*. Estas plataformas permiten a los usuarios interactuar en tiempo real, compartir opiniones y participar en discusiones, lo que también ha generado un aumento en los casos de violencia verbal y ciberacoso.

2.3.3 Análisis de Sentimientos

El análisis de sentimientos, también conocido como minería de opiniones, es una técnica dentro del procesamiento del lenguaje natural (PLN) que se utiliza para identificar, extraer y clasificar emociones, opiniones o actitudes en textos. Esta técnica permite determinar si un texto es positivo, negativo o neutral, y se usa en el contexto de redes sociales para identificar comportamientos agresivos o violentos.

2.3.4 Procesamiento del Lenguaje Natural (PLN)

El procesamiento del lenguaje natural es una rama de la inteligencia artificial (IA) que estudia la interacción entre las computadoras y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten y generen lenguaje humano de manera eficiente. En el contexto de esta tesis, el PLN es fundamental para analizar y clasificar el contenido textual de redes sociales, detectando patrones de violencia verbal.

2.3.5 Detección de Violencia Verbal

La detección de violencia verbal implica identificar automáticamente signos de abuso, agresión o lenguaje dañino en textos. Utiliza técnicas de PLN y análisis de sentimientos para filtrar y clasificar contenidos potencialmente ofensivos en redes sociales. El objetivo es prevenir la propagación de lenguaje violento en línea y proteger a los usuarios de posibles daños emocionales.

2.3.6 Modelos de Aprendizaje Automático

Los modelos de aprendizaje automático *(machine learning)* son algoritmos que permiten a las máquinas aprender patrones a partir de datos. En el contexto de la detección de violencia verbal, se entrenan modelos para identificar automáticamente características del lenguaje violento en grandes volúmenes de datos textuales, como publicaciones en redes sociales.

2.3.7 Precisión en la Detección

La precisión en la detección se refiere a la capacidad de un modelo o sistema para identificar correctamente las instancias de violencia verbal sin generar muchos falsos positivos (cuando el sistema identifica un mensaje no violento como violento). Es un indicador clave de la efectividad de un sistema de análisis de sentimientos aplicado a este tipo de problemas.

2.3.8 Eficiencia Computacional

La eficiencia computacional mide los recursos, como tiempo y poder de procesamiento, que un sistema requiere para realizar una tarea. En el contexto de esta tesis, la eficiencia computacional es importante para procesar grandes cantidades de datos en redes sociales en tiempo real, minimizando el uso de recursos sin comprometer la precisión en la detección de violencia verbal.

2.3.9 Ciberacoso

El ciberacoso es una forma de acoso que se realiza a través de medios digitales, como redes sociales, mensajes de texto o correos electrónicos. Implica el uso de lenguaje abusivo, amenazas, o la difusión de información falsa para

intimidar, controlar o humillar a una persona. La detección de violencia verbal es una herramienta clave para combatir el ciberacoso en línea.

2.3.10 Algoritmos de Clasificación

Los algoritmos de clasificación son un tipo de técnica de aprendizaje automático que asignan categorías a los datos. En esta tesis, estos algoritmos se utilizan para clasificar el contenido textual como violento o no violento, basándose en las características aprendidas de ejemplos anteriores.

Capítulo III: Hipótesis y variables

3.1. Hipótesis

3.1.1. Hipótesis General

El análisis de sentimientos aplicado a comentarios y publicaciones en redes sociales permite identificar de manera precisa y eficiente señales de violencia verbal.

3.1.2. Hipótesis Específicas

- Los comentarios y publicaciones en redes sociales que contienen señales de violencia verbal presentan patrones lingüísticos específicos, como el uso frecuente de palabras agresivas y una alta frecuencia de comentarios con tono agresivo.
- Las técnicas de procesamiento de lenguaje natural (PLN) y análisis de sentimientos, como machine *learning* y *deep learning*, son eficaces para identificar violencia verbal en redes sociales con un alto nivel de precisión y rendimiento.
- El modelo de análisis de sentimientos desarrollado puede alcanzar un nivel significativo de precisión, sensibilidad y especificidad en la detección de violencia verbal en redes sociales.
- Los informes y visualizaciones generados a partir del análisis de sentimientos permiten identificar patrones y tendencias de violencia verbal, y la implementación de alertas automáticas es útil para la detección temprana de comportamientos violentos en redes sociales.

3.2. Operacionalización de variables

3.2.1 Variable 1

Comentarios y publicaciones de redes sociales

Es el contenido textual publicado por los usuarios en la plataforma *Facebook*. Para fines de este estudio, se centran en los comentarios y publicaciones que pueden contener señales de violencia verbal.

3.2.1.1 Dimensiones:

A. Palabras clave relacionadas con violencia verbal:

Definición conceptual. Términos o expresiones comúnmente asociadas con violencia verbal, como insultos, amenazas, descalificaciones, etc.

Definición operacional. Frecuencia de aparición de un conjunto de palabras clave predefinidas (basadas en léxicos de violencia verbal) dentro de los comentarios y publicaciones.

Indicadores: Número de ocurrencias de palabras o frases violentas por comentario o publicación.

B. Frecuencia de comentarios violentos:

Definición conceptual. La cantidad de comentarios o publicaciones que contienen contenido violento, agresivo o amenazante en un periodo de tiempo.

Definición operacional. Conteo de publicaciones y comentarios que cumplen con criterios de agresividad o violencia verbal (definidos por el uso de palabras clave y frases con intencionalidad violenta).

Indicadores. Porcentaje de comentarios agresivos sobre el total de comentarios analizados.

3.2.2 Variable 2

Categoría del sentimiento de violencia verbal.

Es la categoría asociada al sentimiento de violencia verbal en los comentarios y publicaciones en redes sociales.

3.2.2.1 Dimensiones:

A. Nivel de violencia en el texto:

Definición conceptual. Intensidad o severidad de la violencia verbal detectada en el texto, determinada por el tono y las palabras utilizadas.

Definición operacional. Asignación de puntuaciones o etiquetas de agresividad a cada comentario, clasificados en categorías (por ejemplo, leve, moderada, alta) según las características del texto.

Indicadores. Valor numérico de la agresividad detectada en función de un modelo preentrenado (machine *learning* o *deep learning*).

B. Precisión y rendimiento del modelo:

Definición conceptual. Capacidad del modelo de análisis de sentimientos para identificar correctamente comentarios y publicaciones violentas en comparación con comentarios no violentos.

Definición operacional. Cálculo de métricas de clasificación como precisión, sensibilidad (recall), especificidad y f1-score después de aplicar el modelo de análisis de sentimientos a los datos preprocesados.

Indicadores:

- Precisión. Proporción de comentarios identificados correctamente como violentos en relación con todos los comentarios analizados.
- Sensibilidad. Proporción de comentarios violentos que el modelo identifica correctamente sobre el total de comentarios violentos.
- Especificidad. Proporción de comentarios no violentos que el modelo identifica correctamente sobre el total de comentarios no violentos.

3.3. Matriz de operacionalización de variables

Tabla 1

Operacionalización las variables

Variable	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Escala de valoración	Instrumentos
Comentarios y publicaciones de redes sociales.	Es el contenido textual publicado por los usuarios en la plataforma Facebook. Para fines de este estudio, se centran en los comentarios y publicaciones que pueden contener señales de violencia verbal.	Para la medición de la variable comentarios y publicaciones de redes sociales se descompondrá en las dimensiones palabras clave relacionadas con la violencia verbal y la frecuencia de comentarios violentos.	Palabras clave relacionadas con violencia verbal	Número de ocurrencias de palabras o frases violentas por comentario o publicación.	Conteo	Registro de ocurrencias
			Frecuencia de comentarios violentos	Porcentaje de comentarios agresivos sobre el total de comentarios analizados.		
Categoría del sentimiento de violencia verbal.	Es la categoría asociada al sentimiento de violencia verbal en los comentarios y publicaciones en redes sociales.	Para la medición de la variable identificación de violencia verbal mediante análisis de sentimientos se realizará a través de la descomposición de esta en las dimensiones nivel de violencia en el texto, y precisión y rendimiento del modelo.	Nivel de violencia en el texto	Valor numérico de la agresividad detectada en función de un modelo preentrenado (machine *learning* o *deep learning*).	Categoría	Modelo de análisis de sentimiento
			Precisión y rendimiento del modelo	• Precisión: Proporción de comentarios identificados correctamente como violentos en relación con todos los comentarios analizados. • Sensibilidad: Proporción de comentarios violentos que el modelo identifica correctamente sobre el total de comentarios violentos. • Especificidad: Proporción de comentarios no violentos que el modelo identifica		

				correctamente sobre el total de comentarios no violentos.		

Capítulo IV: Metodología del estudio

4.1. Enfoque, tipo y alcance de investigación

4.1.1. Enfoque

El presente estudio adopta un enfoque de investigación mixto, integrando métodos cuantitativos y cualitativos en función de los objetivos planteados. El enfoque cuantitativo se fundamenta en la medición precisa de variables y en el tratamiento estadístico de datos, siguiendo lo descrito por Ñaupas et al. (2018, p. 140), ya que busca analizar y cuantificar características específicas en comentarios de redes sociales que sugieren violencia verbal. A su vez, este enfoque orienta la investigación hacia aplicaciones prácticas, con la finalidad de desarrollar una herramienta digital capaz de medir y clasificar el nivel de violencia verbal en comentarios y publicaciones, facilitando así la identificación de patrones de riesgo.

Por otra parte, Cabrera (2023) destaca que la ciencia de datos, al combinar técnicas de ciencias de la computación y estadística para la recopilación y análisis de grandes volúmenes de datos, también puede considerarse dentro de un enfoque cualitativo con una orientación interdisciplinaria. Este enfoque incluye métodos como el análisis de redes sociales (*netnografía*), el análisis de sentimientos y *big* data (p. 40). En este contexto, el presente estudio se alinea también con el enfoque cualitativo, dado que tiene como objetivo desarrollar un modelo de análisis de sentimientos mediante minería de texto, orientado a reconocer patrones lingüísticos asociados a la violencia verbal.

Por lo tanto, el estudio se enmarca en un enfoque mixto, en el que el análisis cuantitativo se emplea para medir y clasificar los comentarios en redes sociales, mientras que el análisis cualitativo permite una interpretación profunda de los patrones de lenguaje y el contexto sociolingüístico en el que estos comentarios se producen. Esta integración de enfoques fortalece la comprensión de las dinámicas de violencia verbal en redes sociales y facilita el desarrollo de herramientas digitales efectivas.

4.1.2. Tipo y alcance

El presente estudio es de tipo aplicado debido a su orientación a la resolución de un problema práctico, concretamente la identificación de señales de violencia verbal en redes sociales mediante el uso de herramientas de ciencia de datos. Esta característica refleja un propósito dirigido a la generación de soluciones concretas y funcionales.

En cuanto a su alcance, se considera descriptivo, ya que se centra en la recopilación y el análisis de datos relacionados con las características, propiedades, dimensiones y clasificaciones de los comentarios en redes sociales que sugieren violencia verbal. Este enfoque está alineado con lo planteado por Ñaupas et al. (2018, p. 134), quienes destacan la importancia de describir y categorizar fenómenos en contextos específicos. Además, el estudio también posee un alcance explicativo, ya que se orienta a comprender el funcionamiento y efectividad de distintas técnicas y modelos de clasificación en la identificación de violencia verbal en redes sociales. Así, no solo se busca identificar patrones de violencia verbal, sino también explicar cómo y por qué determinados modelos y métricas son más eficaces en este contexto y en qué condiciones se manifiestan estas relaciones (Hernández et al., 2014, p. 95).

4.2. Diseño de la investigación

Un diseño secuencial exploratorio, se inicia con una fase cualitativa que permite explorar y obtener una comprensión profunda del fenómeno en estudio (Cueva et al., 2023, p. 100). En tal sentido, en el presente estudio, la primera fase tiene el objetivo de identificar patrones y características específicas del lenguaje en comentarios y publicaciones, utilizando técnicas de análisis de sentimientos y PLN. A través de este enfoque cualitativo, se pueden clasificar los matices lingüísticos asociados a la violencia verbal, lo cual proporciona una base interpretativa valiosa en detalles sobre el contenido presente en redes sociales.

Luego, se procede a una fase cuantitativa en la cual se desarrolla un modelo de *machine learning*. Este modelo cuantifica la presencia de violencia verbal en los comentarios y permite medir tanto la intensidad como la probabilidad de violencia

en el texto. Al agregar esta precisión cuantitativa, se facilita la generalización de los hallazgos a un conjunto de datos más amplio, logrando así validar y extender los resultados obtenidos en la fase cualitativa inicial. Además, al no manipularse ninguna variable para analizar efectos causales, el estudio posee un diseño no experimental y, debido a su enfoque mixto, también se considera observacional.

4.3. Población y muestra

4.3.1. Población

Una población, o universo, se refiere al conjunto de elementos (como sujetos, objetos o entidades abstractas) que comparten una o más características. Generalmente, el término población implica el conjunto completo de las unidades de análisis que se desea investigar, y se define al especificar las características comunes de estos elementos (Pardo, Ruiz & San Martín, 2012). En ese sentido, la población del presente estudio está constituida por comentarios en la red social *YouTube* en Perú durante el año 2024.

4.3.2. Muestra

Gutiérrez (2016) afirmó que la muestra representativa no debe ser un modelo reducido a la población sino un concepto asociado con las estrategias de muestreo; es decir, la representatividad de una muestra se argumenta en la dupla conformada por el diseño del muestreo y estimador; que permitan estimar exactamente los totales poblacionales (pp. 59-60). Por tal motivo, en el presente estudio la muestra está conformado por 9556 comentarios extraídos de la plataforma *YouTube* que contengan comentarios violentos y no violentos.

4.4. Técnicas e instrumentos de recolección de datos

4.4.1. Técnicas e instrumentos

En este estudio se empleará la Automatización de Procesos Robóticos (RPA) para la recolección de datos, utilizando específicamente *Web Scraping*. Esta técnica es ampliamente utilizada en investigación cuantitativa, ya que permite extraer grandes volúmenes de datos de forma sistemática y eficiente, obteniendo

información precisa y actualizada de plataformas en línea (Van Hoecke, H. y Pauwels, J., 2018).

En este caso, el objetivo principal es recopilar comentarios y publicaciones en Facebook que se relacionen con temas relevantes, como las señales de violencia y los comportamientos observados en redes sociales.

La herramienta de *Web Scraping* se configurará meticulosamente para capturar datos relevantes, aplicando filtros que aseguren la pertinencia de la información con los objetivos de la investigación. Esto incluye la selección de páginas y perfiles públicos específicos, así como la definición de parámetros que limiten la extracción a publicaciones y comentarios con ciertos términos, facilitando un análisis alineado con el propósito del estudio. El diseño de esta estrategia de recopilación se basa en una revisión exhaustiva de estudios sobre análisis de redes sociales y el uso de RPA en la extracción de textos, siguiendo las buenas prácticas éticas y de privacidad.

Asimismo, el preprocesamiento de datos incluirá pasos como limpieza de texto, eliminación de duplicados y normalización, lo que permitirá estructurar la información de forma adecuada para el análisis posterior. Este enfoque sistemático para la extracción y preparación de los datos es esencial para evitar sesgos y captar el contexto completo de los comentarios obtenidos (Crane, 2019).

La elección de RPA y *Web Scraping* como técnica central en este estudio se justifica por su capacidad para ofrecer una visión amplia y detallada de la actividad en redes sociales. (Davenport, T. H. y Kirby, J., 2020) sugieren que las técnicas de automatización como RPA son especialmente adecuadas en investigaciones que requieren grandes volúmenes de datos de fuentes digitales públicas, facilitando el acceso a información que sería difícil de recopilar manualmente. Este enfoque permitirá obtener una visión detallada de cómo se expresan y comparten señales de violencia en redes sociales, proporcionando resultados que reflejen con precisión las dinámicas de interacción en estos entornos digitales.

Figura 3

Código fuente para la recolección de comentarios

```
from selenium import webdriver
from selenium.webdriver.common.by import By
import pandas as pd
from selenium.webdriver.firefox.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.firefox.options import Options
from webdriver_manager.firefox import GeckoDriverManager
import time

options = Options()
options.headless = True
service = Service(GeckoDriverManager().install())
driver = webdriver.Firefox(service=service, options=options)

def youtube(urls):
    comentarios = []  # Initialize an empty list to store comments

    # Assuming `driver` is already initialized before this function is called

    for url in urls:
        driver.get(url)
        time.sleep(5)

        # Scroll and collect comments
        for _ in range(16):
            try:
                WebDriverWait(driver, 5).until(
                    EC.visibility_of_element_located((By.ID, 'comments'))
                )
                time.sleep(3)
                comments = driver.find_elements(By.ID, 'comments')

                if comments:
                    driver.execute_script("window.scrollBy(0, 800);")
            except Exception as e:
                print(f"Error: {e}")
                break

        # Extract the comments
        elementos = driver.find_elements(By.CSS_SELECTOR, 'ytd-comment-thread-
renderer')

        for elemento in elementos:
            try:
                comentario_text = elemento.find_element(By.CSS_SELECTOR, '#content-
text').text
                comentarios.append(comentario_text)  # Add the comment to the list
                print("->: ",comentario_text)
            except Exception as e:
                print(f"No se pudo encontrar el comentario: {e}")

    # Create a DataFrame from the collected comments
    df = pd.DataFrame(comentarios, columns=["Comentario"])

    # Remove duplicates based on the comment text
    dataset = df.drop_duplicates(subset="Comentario")
    # Write the dataset to a CSV file
    dataset.to_csv('Youtubedataset.csv', index=False, encoding='utf-8' , sep = ',')

    #return df
```

4.4.2. Validez y confiabilidad

Para garantizar la validez y confiabilidad en la recolección de datos a través de *Web Scraping*, es fundamental adoptar buenas prácticas y principios de RPA. La validez implica la precisión en la extracción de datos, lo que requiere definir criterios claros y realizar validaciones, como la verificación cruzada con fuentes fidedignas (Kitchin, 2014).

La confiabilidad se relaciona con la consistencia de los resultados. Para asegurarla, se deben realizar pruebas exhaustivas del código y aplicar pruebas unitarias para evaluar su rendimiento en diversos escenarios (Robert, 2011). Además, es esencial manejar excepciones, lo que permite al sistema adaptarse a cambios en la estructura de las páginas sin comprometer la calidad de los datos (Beck, 2001).

Una documentación clara del código es crucial para su mantenimiento y comprensión (Cunningham, 2015). Realizar revisiones de pares ayuda a identificar errores y optimizar el código, mejorando su calidad. Asimismo, el uso de herramientas de seguimiento de versiones facilita el control de cambios, asegurando la consistencia del código (Sussman, G. y Steele, G., 1975). Adoptar estas prácticas no solo mejora la validez y confiabilidad del proceso de *Scraping*, sino que también se alinea con los principios de RPA, asegurando una automatización efectiva y eficiente (Willcocks et al., 2015).

4.4.3. Procedimiento de recolección de datos

La recolección de datos se llevó a cabo mediante la técnica de *Web Scraping* en la plataforma *YouTube*, con el propósito de extraer comentarios relevantes para el análisis de violencia verbal. El proceso inició con la recopilación automatizada de comentarios a partir de 1646 enlaces seleccionados aleatoriamente, abarcando diversos ámbitos sociales como política, deporte, economía, psicología y problemáticas sociales. A diferencia de un enfoque basado en un individuo objetivo, la estrategia implementada se centró en capturar una muestra amplia y representativa a partir de publicaciones y debates públicos en redes sociales, garantizando diversidad en las expresiones lingüísticas y opiniones recogidas.

La elección de *YouTube* como plataforma principal se fundamentó en su amplia base de usuarios y la naturaleza abierta de sus interacciones, lo que permitió acceder a contenido variado para el estudio. Durante la extracción de datos, se aplicaron filtros específicos para descartar emoticones, textos excesivamente cortos, caracteres aleatorios y comentarios de longitud extrema. Luego, se implementó un proceso de clasificación supervisada en el que un lingüista evaluó individualmente cada comentario para determinar su naturaleza violenta o no violenta, asegurando mayor precisión en la categorización de los datos.

En cuanto a la gestión de la información recolectada, se respetaron estrictamente las políticas de privacidad de *YouTube*, garantizando que el acceso y uso de los datos se ajustara a los principios de transparencia y ética en la investigación digital. Para proteger la integridad de la información, se implementaron protocolos de seguridad en el almacenamiento y manejo de los datos, además de un procedimiento estructurado para la eliminación de la información una vez finalizado el análisis.

Este enfoque metodológico permitió obtener una base de datos depurada de 2343 comentarios, categorizados en violentos (1) y no violentos (0), lo que facilitó el desarrollo de modelos de análisis de sentimientos y PLN. Así, la recolección de datos proporcionó una base sólida para la identificación de patrones lingüísticos asociados a la violencia verbal en redes sociales, asegurando un análisis riguroso y alineado con los estándares éticos en la investigación digital.

4.5. Técnicas de análisis de datos

En el presente estudio, cuyo objetivo es Identificar señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos, se optó por técnicas de clasificación basadas en aprendizaje automático. Esta metodología permitió examinar cómo diversas características de publicaciones en redes pueden predecir o clasificar contenido violento, facilitando la identificación temprana de conductas de riesgo. (Manning, C. y Schütze, H., 1999) destacan que los modelos de clasificación en PLN son herramientas poderosas para analizar patrones complejos en textos, posibilitando no solo la

identificación de relaciones entre variables, sino también la cuantificación de su precisión en la detección de violencia.

Para implementar este análisis, se emplearon bibliotecas como *Scikit-Learn*, *TensorFlow* y *PyTorch*, reconocidas en ciencia de datos y aprendizaje profundo. Scikit-Learn es clave en el preprocesamiento de datos y en la creación de modelos de clasificación y regresión, mientras que *TensorFlow* y *PyTorch* son ideales para el desarrollo y entrenamiento de redes neuronales profundas, captando patrones complejos en textos de redes sociales. La selección de estas herramientas permite obtener resultados confiables y generalizables, fundamentales para un estudio que busca proporcionar recomendaciones empíricas para prevenir la violencia en entornos digitales.

Capítulo V: Resultados

5.1 Análisis de los resultados

En el presente capítulo se exponen los hallazgos derivados del proceso de extracción y análisis de comentarios provenientes de redes sociales, con el propósito de identificar y etiquetar aquellos que contienen señales de violencia verbal. Posteriormente, se describe el desarrollo e implementación de un modelo de análisis de sentimientos basado en técnicas de PLN, diseñado para la detección automatizada de expresiones violentas en los textos analizados. A continuación, se presenta la evaluación del desempeño del modelo mediante el uso de métricas de clasificación, con el fin de optimizar su precisión y capacidad predictiva. Finalmente, se introduce el diseño y funcionalidad de una herramienta digital, concebida para generar informes analíticos y visualizaciones dinámicas que permitan identificar patrones de violencia verbal en los comentarios procesados.

5.1.1 Exploración y preprocesamiento de datos:

En el presente estudio titulado *"Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos"*, se diseñó y aplicó un procedimiento estructurado para la exploración y preprocesamiento de los datos, compuesto por diversas etapas metodológicas.

Inicialmente, se realizó la recolección de datos a través de la extracción automatizada de comentarios en las plataformas *Facebook* y *YouTube*. Mediante la técnica de *Web Scraping*, se obtuvieron un total de 9556 comentarios provenientes de 1646 enlaces seleccionados aleatoriamente, abarcando diversos ámbitos sociales, tales como política, deporte, economía, psicología, problemáticas sociales y entretenimiento. El *Web Scraping* permitió la captura sistemática de datos a través de scripts especializados, los cuales accedieron a las *Páginas Web*, extrajeron el contenido relevante y lo almacenaron en un formato estructurado para su posterior análisis.

Posteriormente, se llevó a cabo un primer proceso de filtrado de los comentarios, eliminando aquellos que contenían emoticones, textos excesivamente cortos o atípicos, tales como caracteres aleatorios. También se

presentaron comentarios de longitud extrema. Como resultado de esta etapa, la base de datos se redujo a 4507 comentarios.

A continuación, se realizó un proceso de clasificación semántica bajo la supervisión de un lingüista, analizando cada comentario individualmente para determinar su naturaleza violenta o no violenta. Durante esta fase, se identificaron comentarios ambiguos, es decir, aquellos que contenían términos o expresiones groseras pero cuyo significado variaba según el contexto.

Ejemplo

contigo no solo se aprende historia, se aprende de verdad el estilo de vida de cada lugar que muchos ignoran en el mundo. ¡gracias, Luisito!	0
bien mierda, felicitaciones por tus logros.	0
matutazo, dale (u) toda la vida, a llorar kgones ctmr	1
pagaron un montón de dinero para un espectáculo aburrido y a un ídolo de masas en caída libre.	1

Para garantizar una categorización precisa, estos comentarios fueron excluidos del análisis. Como resultado de este procedimiento, se obtuvo una muestra depurada de 2343 comentarios.

Una vez validada la muestra, se procedió a la construcción de una base de datos estructurada en formato matricial utilizando *Excel*, en la que cada observación se organizó en tres columnas: la primera contenía un código identificador único para cada comentario, la segunda incluía el texto del comentario, y la tercera registraba su clasificación binaria como violento (1) o no violento (0). Para el procesamiento y análisis de los datos, se emplearon diversos recursos computacionales, incluyendo *Google Colab* con capacidad de procesamiento mediante GPU, así como herramientas de manipulación de datos como *pandas*, numpy, *matplotlib*, *seaborn*, *wordcloud*, *stats* y skew.

El tratamiento de los datos incluyó un análisis descriptivo de los comentarios de la muestra final, considerando métricas como la distribución del número de palabras por comentario, la frecuencia de términos violentos dentro de cada comentario y el porcentaje de comentarios clasificados como violentos (1) y no

violentos (0). Asimismo, se realizó una visualización de las palabras agresivas más utilizadas en los comentarios categorizados como violentos.

Finalmente, la identificación de términos violentos se fundamentó en la creación de un diccionario especializado, construido a partir de las palabras más recurrentes en los comentarios etiquetados como agresivos.

5.1.1.1 Análisis descriptivo de los comentarios

a. ***Frecuencias del número de palabras por comentarios***

Para la distribución del número de palabras por comentarios se construyó la Tabla 2, la cual describe la cantidad de palabras contenidas en los comentarios analizados, agrupadas en intervalos definidos, proporcionando información sobre su frecuencia absoluta, relativa y acumulada.

Se observa que la mayoría de los comentarios analizados contienen entre 1 y 33 palabras, representando el 89.54% del total, con 2098 ocurrencias. El siguiente rango más frecuente es el de 33 a 65 palabras, con 186 comentarios (7.94%), lo que indica que más del 97% de los comentarios contienen hasta 65 palabras. Conforme aumenta el número de palabras por comentario, la frecuencia disminuye progresivamente, con solo un 1.49% de comentarios en el intervalo de 65 a 97 palabras y valores marginales en los rangos superiores.

El análisis de la frecuencia relativa acumulada evidencia que el 99.91% de los comentarios tienen menos de 321 palabras, lo que sugiere que los comentarios en redes sociales, en su gran mayoría, son concisos y de corta extensión. Solo un porcentaje mínimo (0.09%) de los comentarios analizados supera las 321 palabras, evidenciado por la baja frecuencia en los últimos intervalos.

Estos resultados son fundamentales para comprender la estructura del texto en comentarios de redes sociales y pueden influir en el desarrollo de modelos de análisis de lenguaje natural.

Tabla 2

Frecuencias del número de palabras por comentarios

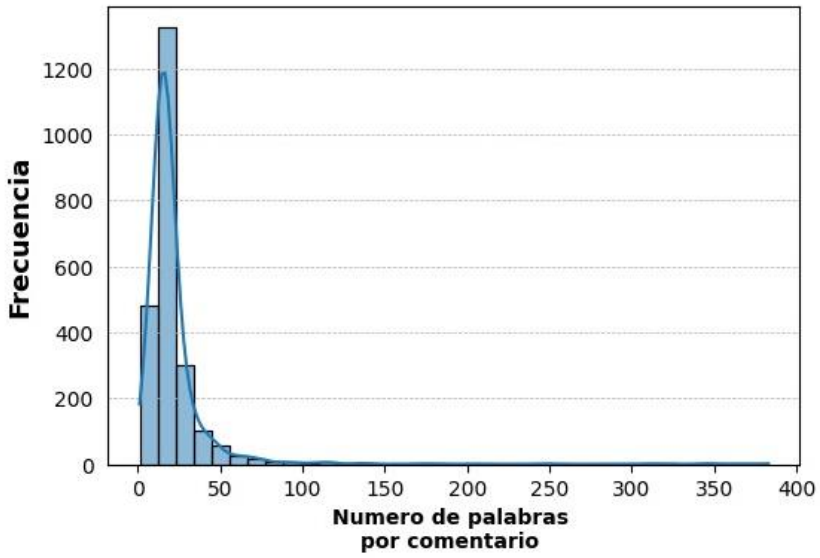
Numero palabras	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada
[1, 33)	1	33	2098	89.54%	2098	89.54%
[33, 65)	33	65	186	7.94%	2284	97.48%
[65, 97)	65	97	35	1.49%	2319	98.98%
[97, 129)	97	129	11	0.47%	2330	99.45%
[129, 161)	129	161	3	0.13%	2333	99.57%
[161, 193)	161	193	2	0.09%	2335	99.66%
[193, 225)	193	225	1	0.04%	2336	99.70%
[225, 257)	225	257	1	0.04%	2337	99.74%
[257, 289)	257	289	0	0.00%	2337	99.74%
[289, 321)	289	321	2	0.09%	2339	99.83%
[321, 353)	321	353	2	0.09%	2341	99.91%
[353, 385)	353	385	2	0.09%	2343	100.00%

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

En la figura 4, se observa la predominancia de comentarios breves, lo cual sugiere que las estrategias de preprocesamiento y modelado deben adaptarse a textos de extensión limitada, optimizando técnicas de *tokenización* y representaciones vectoriales.

Figura 4

Distribución del número de palabras por comentario



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

La Tabla 3 presenta las medidas de resumen de la distribución del número de palabras por comentario. Estos estadísticos descriptivos proporcionan información clave sobre la dispersión, centralización y forma de la distribución de palabras en los comentarios analizados.

El total de observaciones es de 2343 comentarios, con una media de 20.79 palabras por comentario, lo que indica que, en promedio, los comentarios analizados son relativamente cortos. Sin embargo, la desviación estándar (22.94) sugiere una considerable variabilidad en la extensión de los comentarios. La longitud mínima registrada es de 1 palabra, mientras que la máxima alcanza 383 palabras, evidenciando una gran asimetría en la distribución de los datos.

El análisis de los percentiles muestra que el 25% de los comentarios tienen 12 palabras o menos, el 50% (mediana) contienen hasta 17 palabras, y el 75% presentan 22 palabras o menos, confirmando que la mayoría de los comentarios son breves y están concentrados en un rango reducido de palabras.

Los valores de *kurtosis* (113.47) y *sesgo* (8.98) indican una distribución altamente sesgada hacia la derecha con una marcada concentración de datos en valores bajos y la presencia de valores extremos o *outliers* en el extremo superior de la distribución. Esta alta curtosis sugiere que la mayoría de los comentarios tienen una extensión mucho menor que el valor máximo registrado, lo que se traduce en una distribución leptocúrtica con una cola pesada.

Estos resultados son cruciales para el modelado y preprocesamiento de datos en el análisis de lenguaje natural, ya que reflejan la necesidad de técnicas adecuadas para manejar la alta dispersión y asimetría en la longitud de los comentarios, lo que puede influir en la selección de modelos de clasificación y estrategias de normalización en el análisis de sentimientos y detección de violencia en redes sociales.

Tabla 3

Medidas de resumen de la distribución del número de palabras por comentario

Medidas de resumen	Estadístico
count	2343
mean	20.79257
std	22.94034
min	1
25%	12
50%	17
75%	22
max	383
kurtosis	113.4691
skew	8.987774

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

b. Frecuencias del número de palabras agresivas por comentario clasificado como violento.

En la Tabla 4 se presenta la distribución de la frecuencia de términos violentos dentro de cada comentario clasificado como violento; el cual permite examinar la cantidad de términos agresivos o violentos presentes en los comentarios identificados como violentos, proporcionando una perspectiva detallada sobre la densidad de lenguaje hostil en los textos analizados.

Los resultados indican que la mayoría de los comentarios violentos contienen entre 1 y 33 términos violentos, representando un 89.54% del total. Esto sugiere que la violencia verbal en redes sociales tiende a manifestarse a través de mensajes cortos que contienen una alta carga de agresividad en pocas palabras. En contraste, solo el 7.94% de los comentarios violentos contienen entre 33 y 65 términos violentos, mientras que aquellos con más de 65 términos violentos representan un porcentaje marginal de la muestra total.

A medida que aumenta la cantidad de términos violentos por comentario, la frecuencia relativa disminuye de manera significativa, con valores inferiores al 2% en los intervalos superiores a 65 términos violentos. La distribución acumulada muestra que el 98.98% de los comentarios violentos contienen menos de 97

términos agresivos, lo que indica una fuerte concentración de lenguaje violento en comentarios relativamente cortos.

Tabla 4

Distribución de la frecuencia de términos violentos dentro de cada comentario violentos

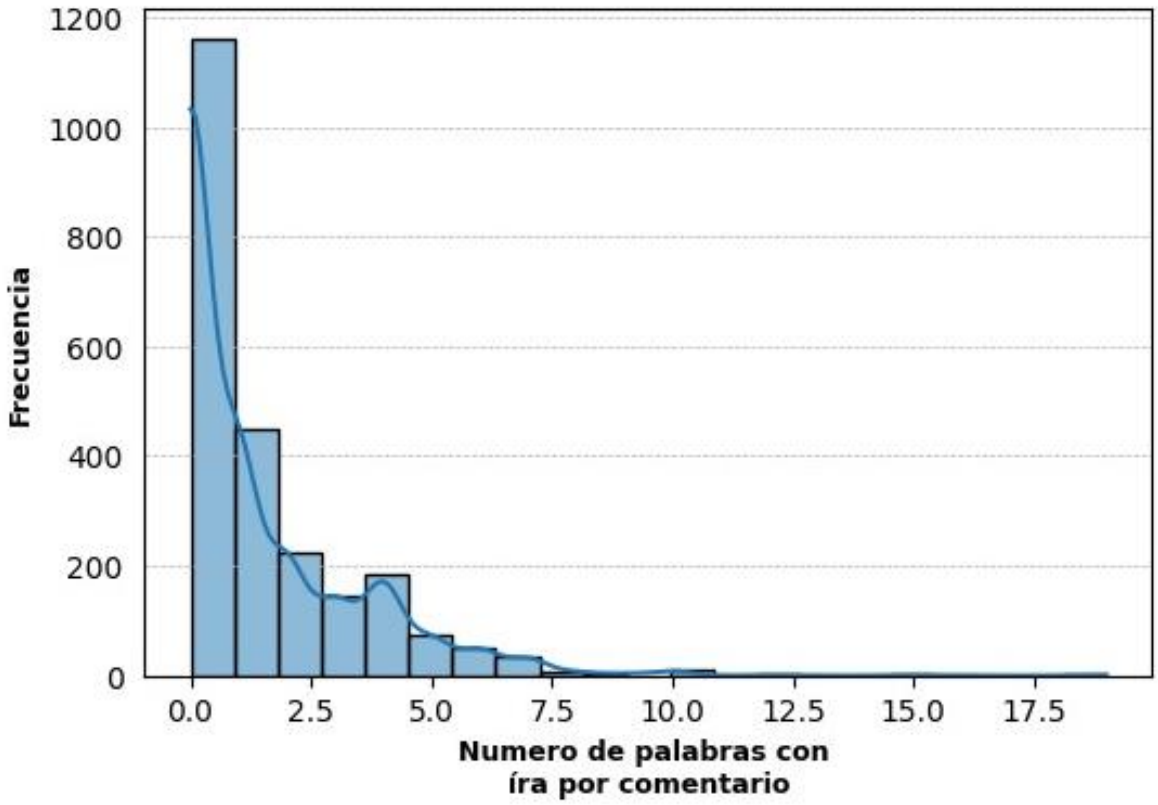
Numero palabras	Límite Inferior	Límite Superior	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada
[1, 33)	1	33	2098	89.54%	2098	89.54%
[33, 65)	33	65	186	7.94%	2284	97.48%
[65, 97)	65	97	35	1.49%	2319	98.98%
[97, 129)	97	129	11	0.47%	2330	99.45%
[129, 161)	129	161	3	0.13%	2333	99.57%
[161, 193)	161	193	2	0.09%	2335	99.66%
[193, 225)	193	225	1	0.04%	2336	99.70%
[225, 257)	225	257	1	0.04%	2337	99.74%
[257, 289)	257	289	0	0.00%	2337	99.74%
[289, 321)	289	321	2	0.09%	2339	99.83%
[321, 353)	321	353	2	0.09%	2341	99.91%
[353, 385)	353	385	2	0.09%	2343	100.00%

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

En la figura 5, se resaltan estos hallazgos ya que son de suma importancia para el análisis de violencia verbal en redes sociales, ya que permiten comprender la extensión típica de los comentarios agresivos y su posible impacto en los usuarios. La alta concentración de términos violentos en comentarios breves sugiere la necesidad de enfoques de detección eficientes que sean capaces de identificar señales de violencia en textos de corta longitud. Además, la baja frecuencia de comentarios con un alto número de términos violentos podría indicar la presencia de discursos más elaborados que requieren técnicas avanzadas de análisis semántico para su correcta clasificación.

Figura 5

Distribución de la frecuencia de términos violentos dentro de cada comentario violentos



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

La Tabla 5, presenta las medidas de resumen correspondientes a la distribución del número de palabras violentas por comentario clasificado como agresivo, dentro del estudio con el fin de comprender la variabilidad y la tendencia central de la cantidad de términos violentos presentes en los comentarios identificados como agresivos.

Los resultados indican que la muestra está conformada por 2343 comentarios violentos. La media del número de palabras violentas por comentario es de 1.37, lo que sugiere que, en promedio, los comentarios clasificados como agresivos contienen aproximadamente una palabra violenta. Sin embargo, la desviación estándar de 1.95 indica una notable dispersión en los datos, lo que

implica que algunos comentarios pueden contener una cantidad significativamente mayor de términos violentos.

En cuanto a la distribución, se observa que el mínimo de palabras violentas en un comentario es 0, lo que sugiere que algunos comentarios, a pesar de haber sido clasificados como agresivos, no necesariamente contienen términos explícitamente violentos, sino que su contexto o estructura pudo haber influido en su clasificación. El primer cuartil (Q1 = 0) y la mediana (Q2 = 1) refuerzan esta observación, indicando que al menos la mitad de los comentarios contienen una o ninguna palabra violenta. Mientras tanto, el tercer cuartil (Q3 = 2) muestra que el 75% de los comentarios contienen hasta 2 palabras violentas.

El valor máximo observado en la muestra es de 19 palabras violentas en un solo comentario, lo que indica la presencia de mensajes altamente agresivos en la base de datos. La curtosis es extremadamente alta (113.47), lo que sugiere la existencia de valores atípicos que concentran un alto número de términos violentos en pocos comentarios. Adicionalmente, el coeficiente de asimetría (skew = 8.98) revela una distribución altamente sesgada hacia la derecha, lo que indica que la mayoría de los comentarios contienen pocas palabras violentas, mientras que una minoría de ellos presenta una cantidad considerablemente mayor.

Tabla 5

Medidas de resumen de la distribución del número de palabras violentas por comentario clasificado como agresivo.

Medidas de resumen	Estadístico
count	2343
mean	1.367904
std	1.952561
min	0
25%	0
50%	1
75%	2
max	19
kurtosis	113.4691
skew	8.987774

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

En este sentido, estos resultados resaltan la naturaleza heterogénea de la violencia verbal en redes sociales, evidenciando que la mayoría de los comentarios agresivos contienen solo una o dos palabras violentas, mientras que algunos casos excepcionales presentan una mayor densidad de lenguaje agresivo.

c. Proporción de comentarios clasificados como violentos (1) y no violentos (0).

La Tabla 6, muestra la distribución de los comentarios clasificados como violentos (1) y no violentos (0) dentro del estudio *"Detección de señales de violencia en redes sociales: un* enfoque desde la ciencia de datos". Esta clasificación es fundamental para comprender la proporción de contenido potencialmente agresivo en las plataformas analizadas.

Los resultados indican que 1202 comentarios (50.87%) fueron clasificados como no violentos, lo que representa una ligera mayoría en la muestra analizada. Por otro lado, 1161 comentarios (49.13%) fueron identificados como violentos, evidenciando una proporción considerable de publicaciones que contienen algún tipo de agresión verbal.

Tabla 6

Proporción de comentarios clasificados como violentos (1) y no violentos (0).

	Frecuencia	**%**
No violentas	1202	50.87
Violentas	1161	49.13

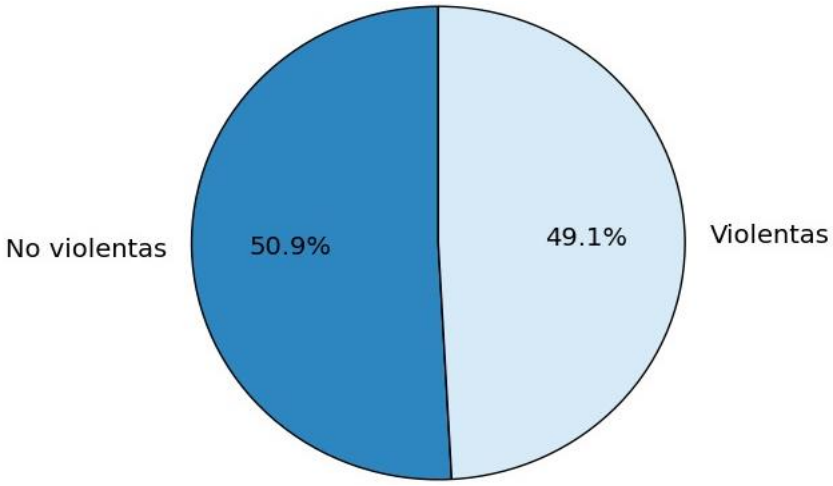
Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

Con el hallazgo expuesto en la Figura 6, se resalta la relevancia de abordar la detección de violencia en redes sociales con enfoques que combinen técnicas de análisis de texto y procesamiento de lenguaje natural, dado que casi la mitad de los comentarios revisados contenían elementos que podrían ser considerados como violentos. La diferencia mínima entre ambas categorías sugiere que la

interacción en redes sociales presenta un alto grado de contenido agresivo, lo que enfatiza la importancia de desarrollar herramientas automatizadas para su monitoreo y análisis.

Figura 6

Distribución de la frecuencia de términos violentos dentro de cada comentario violentos.



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

d. Visualización de las palabras violentas o agresivas con mayor frecuencia.

La imagen presentada en la Figura 7, corresponde a una nube de palabras, una técnica de visualización utilizada en el análisis de texto para representar gráficamente la frecuencia de aparición de términos dentro de un corpus de datos. En este caso, la nube de palabras refleja los términos más recurrentes en los comentarios clasificados como violentos o agresivos dentro del estudio *"Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos"*.

Las palabras que aparecen con mayor tamaño en la nube, como "corrupto", "delincuentes", "puta (pt)", "vergüenza" y "criminales", indican que estos términos tienen una alta frecuencia de aparición en los comentarios analizados. La prevalencia de estas palabras sugiere que los discursos agresivos en redes sociales suelen estar vinculados a temáticas relacionadas con corrupción,

delincuencia y descalificaciones personales. Asimismo, se observan términos ofensivos y despectivos que refuerzan el tono violento de los mensajes recopilados.

El uso de este tipo de visualización es crucial dentro del análisis de sentimientos y procesamiento de lenguaje natural, ya que permite identificar patrones de lenguaje agresivo y focalizar el estudio en términos clave que pueden indicar señales de violencia verbal. Además, esta representación facilita la comprensión del fenómeno en estudio, contribuyendo a la detección y eventual mitigación de discursos de odio en entornos digitales

Figura 7

Nube de las palabras violentas más frecuentes en los comentarios de la misma índole

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

e. Matriz de palabras violentas.

A continuación, se presenta la matriz de palabras violentas (Tabla 7), la cual ha servido como referencia fundamental para la extracción de los resultados descriptivos previamente expuestos. Dicha matriz está compuesta por un total de 462 términos o expresiones de carácter agresivo o violento, utilizados como base para el análisis.

Tabla 7

Matriz de palabras violentas

abandonado	cagada	embrutecido	huevón	lunar	panzón	resbalosa
abominable	cagón	empalagoso	huevona	lunático	panzona de mrd	resentido
abortar	cagona	emputado	huevones	machista	papanatas	retorcido
absurdo	cagonazo	endemoniado	humillante	mala	parásito	ridicula
aburrido	cague	enfermizo	hundido	malandrín	parrasito	ridiculo
abusivo	calaña	enfermo	hurón	maldición	patán	ridículo
acomplejada	calavera	engañador	idiota	maldita	patético	risible
acomplejado	calientacabezas	engreído	idiotas	maldito	pavaso	robán
acuchillado	callejero	enloquecido	idioteces	maleducado	payasada	robar
adicto	calumniador	entrometido	idiotez	malgastador	payaso	roben
afanador	camorrero	envilecido	idiotizado	malhechor	pedante	rotada
afilado	campa	escandaloso	ignorante	malhumorado	pedorro	ruin
afligido	canalla	esclavo	ignorantes	maligno	pegalón	s1m10s
agresivo	canijo	escoria	iletrado	malnacido	pégalos	sabandija
aguafiestas	caradura	esmirriado	iluso	malsano	peleando	sabandoso
ahogado	carajo	espantapájaros	imbécil	maltratador	pelele	sabelotodo
ahumado	carcamal	especial	imbéciles	malvado	pelmazo	sabiondo
alocado	cascarrabias	espeluznante	impostor	mamadas	pelón	sádico
alucinado	castigador	esquizofrénico	improvisado	mandón	pene	salvaje
amargado	cenutrio	estafador	imprudente	mangante	peor	sangrón
ambicioso	cerdo	estéril	impulsivo	maniático	perdedor	sanguinario
analfabeto	chabacano	estiércol	impuro	manipulador	perezoso	sapo
anclado	chamo	estupida	inadaptado	mantenida	perjudicial	sarasa
andrajoso	chancrosa	estupideces	inbecil	mantenido	perra	sarnoso
animal	charlatán	estupidez	incapaz	mañoso	perreando	sátrapa
aniquilado	chato	estupido	incivilizado	maricon	perritas	saudade
anodino	chiflado	estúpido	incompetente	marioneta	perro	seboso
ansioso	chiquilicuatre	evadido	inconsciente	marrajo	perruchas	segundón
apagado	chismoso	exasperante	inconstante	marrana	perturbador	sepulcral
apaleado	chitón	excluido	inculto	matador	perverso	serpiente
apestosa	chivo	execrable	indecente	matar	pesado	serrana
apestoso	chocante	explotador	indeseable	maton	pesetero	serranito
apocado	chola	extorcionador	indiferente	matón	pesimo	serrano
apurado	cholo	facha	indigno	mediocre	pestilente	simplón
argolleros	chucha	facineroso	indios	melancólico	petulante	sin neuronas
arrastrado	chulon	falaz	infame	mentecato	picapleitos	sin vergüenza
arrebatado	chupapene	fallido	infantil	mentirosa	pícaro	siniestro
arrogante	chupapinga	fanático	inferior	mentiroso	pillo	sinsentido

arrugas	cínica	fantoche	infiel	metiche	pinche	soberbio
asaltante	cizañero	farfullero	Infierno	mezquino	pingajo	socarrón
asesina	cobarde	fastidioso	infumable	micropenes	pingüina	soez
asesino	cocainómano	fea	inhumano	miedoso	pingüino	sórdido
asfixiado	cojo	felón	insensato	mierda	pinocha	soso
asno	cojudo	femiburras	insignificante	miserable	piojosa	sucio
asquerosa	colérico	feo	insolente	mitomana	piojoso	sufridor
asqueroso	comadreja	feroz	insufrible	mocho	pirado	sujetavelas
atolondrado	comepinga	fiasco	insulso	mocoso	piraña	sumiso
atontado	comeverga	fingido	insustancial	molesto	pirañas	superficial
atormentado	compinche	flaco	intolerante	monarca	piromana	suspicaz
atrevido	concha tu madre	flojo	inútil	monstruo	pito chico	susurrante
ausente	concha tu vida	follonero	irrecuperable	monstruoso	pizpireto	tacaño
autista	conchuda	forajido	irresponsable	morboso	placera	tarada
avaro	condenado	fraudulento	irritado	morir	plaga	tarado
avergonzado	conera	fregada	irritante	mostra	plañidero	tartamudo
baboso	conero	fregado	jactancioso	mrd	plasta	tedioso
bacín	cornudo	fríos	jadeante	muerta de hambre	plomo	temerario
badulaque	corrupto	frívolo	jamarra	muerto de hambre	pobretona	terco
bailarín	costroso	frustrado	jarra	muertos	pocilga	terruco
bajuno	cpp	fuji rata	jeropas	mugriento	podrida	testarudo
baldado	cretino	fullero	jeta	mugroso	podrido	testículo
baldragas	criminales	fumón de mrd	jodidos	mulo	pollino	tirano
bandido	cruel	furibundo	joyita	murciélago	pomposo	tmr
barato	csmr	furioso	judas	música de mierda	ponzoñoso	tombos
barbaján	ctmr	gamberro	jumento	mustio	porqueria	torpe
basofia	ctmre	gampi	jurásico	mutilado	porquerías	tosco
bastarda	ctv	ganapán	justiciero	narcoestados	prepotente	tragar
bastardo	cuca	gañán	kacheras	narizón	presa	traidor
basura	culo	garboso	kamikaze	narizona	presumido	travieso
batracio	culón	garrapata	karateka	necio	pretencioso	triste
bazofia	cursi	garrapatas	kche	nefando	profanador	tullido
bebita	cutre	garrulo	kinki	nefasto	provocador	turbio
bellaco	dañino	gil	kriollo	negligente	psicópata	ultrajante
bestia	delincuente	gilipollas	kuka	negro	pt	usurero
bicho	demente	glotón	la ptm	negro asqueroso	pudra	usurpadora
bobo	desagradable	gorda	lacras	nervioso	puerco	vacilón
bocazas	desalmado	gordinflón	ladrón	niñato	pueril	vago
bochorno	descarado	gordofóbica	ladrones	no sirven	pugnaz	vampiro
bolas tristes	desdichado	gorrón	laica	nocivo	pulguiento	vanidoso

boludo	desecho	gringo	laika	nómada	puñalera	vejestorio
borracha	desgraciado	grosero	lambebichos	nulo	puñetero	venenoso
borracho	desleal	gruñón	lame bicho	ñangon	pura hevada	verga
botarate	despreciable	guarro	lamebotas	obcecado	pusilánime	vergonzoso
bravucón	desquiciado	guerrillero	lameculos	obeso	puta	vergüenza
bruja	desubicado	gusano	lamentable	obstinado	puto	vieja
bruta	detestable	hablador	lánguido	odian	putrefacto	vil
bruto	diabólico	haragán	largas	odioso	queca	violador
buena para nada	dictador	harapiento	lento	ofensivo	quejica	violín
buitre	dificultoso	hastiado	leproso	ojalá te violen	quejones	viperino
burdo	díscolo	hazmerreír	lerdo	olvidado	quejumbroso	voraz
burguesito	disparate	hediondo	letárgico	opaco	quimérico	vulgar
burra	doble cara	hipócrita	libertino	oportunista	quita la vida	wato
burro	doble moral	hiriente	licencioso	orgulloso	quitamaridos	webada
buscón	dogmático	holgazán	limitado	osado	rancio	xenófobo
cabestro	drástico	horrendo	limosnero	p3n3	rastrero	yerto
cabro	drogado	horrible	llorón	pachorra	rata	zafio
cabrón	drogo	horror	loca	pacífico	ratas	zángano
cacatúa	dudoso	hostil	loco	pajarraco	ratero	zarrapastrosa
cachao	egocéntrico	hu3v4d4z	locos	pajillero	rebelde	zarrapastroso
cachuda	egoísta	huachafon	locuaz	paleto	rebuscado	zorra
cadáver	ejecutor	huebadas	logrero	palurdo	repelente	zorro
cafre	embaucador	hueco	lombriz	paniquea	repulsivo	zote

5.1.2 Modelo de análisis de sentimientos para la detección de violencia verbal en redes sociales.

En esta sección, se presentan los hallazgos clave del análisis de los resultados relacionados con la detección de señales de violencia verbal en redes sociales, utilizando el modelo *DistilBERT*. Para ello, se introduce una matriz de evaluación que combina información textual, clasificación de comentarios y su representación *tokenizada*, lo cual permite un análisis estructurado y eficaz.

5.1.2.1 Introducción al modelo DistilBERT

DistilBERT es una versión optimizada y más ligera del modelo *BERT (Bidirectional Encoder Representations from Transformers),* desarrollado por *Sanh* et al. (2019). Fue presentado por la organización Hugging Face como una alternativa más eficiente para aplicaciones de PLN, conservando el 97% del

rendimiento de *BERT* mientras reduce su tamaño en un 40% y acelera su velocidad de entrenamiento y evaluación. *DistilBERT* se basa en la arquitectura de transformadores, donde los mecanismos de atención bidireccional permiten capturar relaciones contextuales entre palabras dentro de un texto, independientemente de su posición (Vaswani et al., 2017).

Su principal aplicación es en tareas de clasificación de texto, análisis de sentimientos, traducción automática y respuesta a preguntas. En este estudio, *DistilBERT* se emplea para la detección de señales de violencia verbal en comentarios de redes sociales debido a su eficiencia y alta precisión en problemas de clasificación binaria.

5.1.2.2 Definición de la Matriz de Evaluación

La matriz de evaluación presentada en la Tabla 8, ofrece un análisis detallado de los comentarios procesados, abarcando su contenido textual, su clasificación y su representación *tokenizada*, aspectos fundamentales para su posterior análisis computacional. Cada comentario cuenta con un identificador único registrado en la columna "Id", lo que facilita su trazabilidad dentro del conjunto de datos.

Los comentarios reflejan expresiones comúnmente utilizadas en debates y redes sociales, algunas de ellas con una carga semántica negativa o crítica. Para su clasificación, se ha asignado una etiqueta a cada comentario: el valor 0 indica una expresión no violenta, mientras que el valor 1 sugiere un tono agresivo o con potencial de violencia verbal.

Además, cada comentario ha sido transformado mediante un modelo de PLN, obteniendo una representación *tokenizada*. Esta representación se compone de una lista de identificadores numéricos (*input_ids),* donde cada número corresponde a un token dentro del vocabulario del modelo. Asimismo, se incluye una máscara de atención *(attention_mask),* la cual determina qué partes del texto son relevantes para el modelo y cuáles corresponden a información de relleno.

Entre los comentarios analizados se encuentran ejemplos como "Un Maestro se le extraña", clasificado con una etiqueta de 0, lo que sugiere que no contiene elementos ofensivos. En contraste, el comentario "Eres un huevón." recibió la

etiqueta 1, lo que indica un tono potencialmente agresivo o insultante. De manera similar, "Fuera mierda" ha sido etiquetado con 1, reflejando una connotación violenta. Por otro lado, el comentario "Ni el VAR nos salvaría de este nivel de juego." fue clasificado con 0, lo que indica que su contenido es neutral o no ofensivo.

Tabla 8

Análisis manual y tokenización de comentarios clasificados

Id	Comentario	Etiqueta	Tokenizado
1863	Un Maestro se le extraña	0	{'input_ids': tensor([[101, 4895, 25270, 7367, 3393, 4469, 2532, 102]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1]])}
1853	Eres un huevon.	1	{'input_ids': tensor([[101, 9413, 2229, 4895, 20639, 17789, 1012, 102]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1]])}
1859	Fuera mierda	1	{'input_ids': tensor([[101, 11865, 6906, 2771, 2121, 2850, 102]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1]])}
1830	Ni el VAR nos salvaría de este nivel de juego.	0	{'input_ids': tensor([[101, 9152, 3449, 13075, 16839, 16183, 10755, 2401, 2139, 28517, 9152, 15985, 2139, 18414, 20265, 1012, 102]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])}

En conjunto, la tabla ilustra el proceso de conversión del lenguaje natural en datos numéricos, lo cual es un paso esencial para la automatización del análisis de texto en tareas de clasificación de sentimientos y detección de contenido ofensivo. La integración de la información textual original con los resultados de clasificación y su transformación *tokenizada* proporciona una estructura robusta para el análisis, facilitando el desarrollo de modelos avanzados de PLN para la identificación de patrones lingüísticos en redes sociales y otros entornos digitales.

La Figura 8, ilustra el proceso de *tokenización* y codificación en el análisis de sentimientos, un procedimiento esencial dentro del PLN. Este proceso permitió transformar los textos originales en representaciones numéricas que pueden ser interpretadas por modelos de aprendizaje automático para clasificar comentarios en categorías como violentos o no violentos.

El flujo de trabajo comienza con el texto original, compuesto por los comentarios extraídos de plataformas digitales como *YouTube*. Estos textos, en su forma natural, contienen estructuras lingüísticas complejas que requieren un preprocesamiento antes de ser utilizados en modelos de análisis. La primera etapa del proceso es la *tokenización*, que consiste en dividir el texto en unidades más pequeñas llamadas tokens. Dependiendo del enfoque utilizado, estos tokens pueden ser palabras individuales, frases cortas o incluso caracteres específicos. La segmentación en tokens facilita el análisis del contenido textual y permite que los modelos puedan trabajar con fragmentos estructurados del lenguaje.

Posteriormente, los tokens generados deben convertirse en valores numéricos para que puedan ser procesados por algoritmos de aprendizaje automático. En el gráfico se destacan tres métodos de representación numérica del lenguaje, cada uno con enfoques distintos pero complementarios. El primero es TF-IDF *(Term Frequency Inverse Document Frequency),* una técnica estadística que mide la importancia de cada palabra dentro de un conjunto de documentos, ponderando su frecuencia de aparición y relevancia dentro del corpus analizado. Luego se presenta Word2Vec, un modelo basado en redes neuronales que transforma palabras en vectores de alta dimensionalidad, capturando relaciones semánticas y contextuales entre términos. Finalmente, se encuentra BERT *(Bidirectional Encoder Representations from Transformers),* un modelo avanzado que analiza el contexto bidireccional de las palabras en una oración, permitiendo una comprensión más profunda del significado de los textos.

Una vez que los tokens han sido codificados en representaciones numéricas, los datos se introducen en un modelo de clasificación. Este modelo tiene como objetivo principal categorizar los textos en diferentes clases, en este caso, diferenciando entre comentarios violentos y no violentos. La salida del modelo es procesada en la etapa final, correspondiente al análisis de sentimientos,

donde se interpreta el significado de los comentarios y se determina su carga emocional. Esta última fase es clave en el monitoreo automatizado de contenido en redes sociales, ya que permite identificar discursos de odio y agresión de manera eficiente.

En relación a lo anterior, este tipo de sistemas es esencial en el desarrollo de herramientas de moderación de contenido en redes sociales, ayudando a mitigar la propagación de discursos agresivos y proporcionando información relevante para el estudio del comportamiento en línea.

Figura 8

Proceso de tokenización y codificación en análisis de sentimientos.

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

5.1.2.2 Implementación del *Modelo DistilBERT*

El desarrollo de modelos de inteligencia artificial aplicados al procesamiento del lenguaje natural ha permitido abordar problemas complejos en la comunicación digital. En este contexto, la implementación de *DistilBERT* en la detección de violencia en redes sociales representa un avance significativo en la identificación

automática de discursos agresivos o violentos. A través de un proceso estructurado, el modelo es capaz de analizar, procesar y clasificar comentarios, diferenciando aquellos que contienen violencia verbal de los que no (Wolf et al., 2020).

La imagen presentada ilustra este proceso en seis etapas fundamentales, cada una de las cuales desempeña un papel crucial en la construcción de un sistema eficiente de clasificación de texto. A continuación, se detalla el funcionamiento de cada paso en la implementación del modelo.

Figura 9

Procesos para la implementación del Modelo DistilBERT

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

Pasos principales:

a. Carga de Datos

Todo modelo de aprendizaje profundo requiere una fuente confiable de datos para entrenarse y realizar predicciones de manera precisa. En este caso, el proceso comenzó con la recolección y carga de datos textuales provenientes de redes sociales, tales como *Facebook* o *YouTube*.

Estos datos se almacenaron en formatos estándar como **csv o xlsx**, permitiendo una fácil manipulación y estructuración. Durante esta etapa, se verifica que los textos contengan las etiquetas adecuadas para su clasificación:

- **"0" para no violento**
- **"1" para violento**

Se aseguró la integridad y diversidad del conjunto de datos, ya que un *dataset* sesgado o incompleto podría afectar la capacidad del modelo para identificar correctamente expresiones agresivas en diferentes contextos.

b. Limpieza de datos: Preparando el lenguaje para el modelo

El lenguaje empleado en redes sociales es sumamente dinámico y, en muchas ocasiones, informal. Los textos pueden contener errores ortográficos, emojis, caracteres especiales, abreviaturas y diferentes formas de expresión, lo que puede dificultar la interpretación por parte del modelo.

Para garantizar que el modelo procese correctamente el lenguaje, se aplicó una serie de transformaciones al texto original.

El estudio inició con la extracción automatizada de comentarios en *Facebook* y *YouTube,* utilizando la técnica de *Web Scraping.* Se recolectaron 9556 comentarios de 1646 enlaces seleccionados aleatoriamente, abarcando distintos ámbitos sociales.

Tras la recopilación, se realizó un primer filtrado eliminando comentarios con emoticones, caracteres aleatorios o textos extremadamente cortos o largos, reduciendo la base de datos a 4507 comentarios. Luego, un lingüista supervisó la clasificación semántica, identificando comentarios ambiguos, los cuales fueron excluidos para mejorar la precisión del análisis. Esto dejó una muestra final de 2343 comentarios, categorizados en violentos (1) y no violentos (0).

Los datos fueron organizados en una base matricial en *Excel*, con un identificador único, el texto del comentario y su respectiva clasificación.

c. *Tokenización* y Codificación

Una vez que el texto fue limpiado, se procede a convertirlo en un formato numérico que *DistilBERT* pueda procesar. Para ello, se utiliza un *tokenizador*, una herramienta que divide el texto en tokens individuales y los asocia a un índice en el vocabulario del modelo.

DistilBERT tiene su propio *tokenizador preentrenado*, que convierte cada palabra o subpalabra en un vector numérico de dimensión fija. Esto permite al modelo capturar no solo el significado de las palabras, sino también su relación con otras en la oración. Este paso es esencial, ya que garantiza que el modelo pueda interpretar y aprender de la estructura del lenguaje.

d. División en Conjuntos

Para evaluar el rendimiento del modelo de manera objetiva, es necesario dividir los datos en tres subconjuntos (Pedregosa et al., 2011):

d.1 Conjunto de Entrenamiento: Contiene la mayor parte de los datos y se utiliza para que el modelo aprenda los patrones lingüísticos.

d.2 Conjunto de Validación: Permite ajustar hiperparámetros y evitar el sobreajuste, asegurando que el modelo generalice correctamente.

d.3 Conjunto de Prueba: Se usa exclusivamente para medir el rendimiento final del modelo con datos nunca antes vistos.

La correcta distribución de los datos en estas categorías fue clave para garantizar la fiabilidad de las predicciones y evitar sesgos que puedan afectar el desempeño del sistema.

e. Entrenamiento del Modelo

Una vez que los datos han sido preprocesados y divididos en los conjuntos correspondientes, se inicia el entrenamiento del modelo. *DistilBERT*, al ser un modelo preentrenado en grandes volúmenes de datos generales, requiere una fase

de fine-tuning, donde se ajustan sus pesos para la tarea específica de detección de violencia en redes sociales.

Durante esta fase, el modelo analizó 2343 ejemplos de texto etiquetado y convertidos en tensores de *PyTorch* para su procesamiento, ajustando sus parámetros internos para identificar patrones lingüísticos asociados a la violencia verbal. Se optimizaron los elementos clave como:

- El número de épocas (iteraciones sobre el conjunto de datos).
- La tasa de aprendizaje, que regula la velocidad con la que el modelo ajusta sus pesos.
- El tamaño del lote, que define cuántos ejemplos se procesa en cada iteración.

A medida que el modelo avanzó en el entrenamiento, mejoró su capacidad para distinguir entre comentarios agresivos y no violentos, aumentando así su precisión en la clasificación.

f. Evaluación del Modelo:

Una vez finalizado el entrenamiento, se evalúo el desempeño del modelo utilizando diferentes métricas, tales como:

- Precisión (accuracy). Indica el porcentaje de predicciones correctas.
- Recall. Mide cuántos comentarios violentos fueron detectados correctamente.
- F1-score, Calcula un balance entre precisión y recall.

Además, se genera una matriz de confusión, que proporcionó una visión detallada del desempeño del modelo, mostrando cuántas predicciones fueron correctas y cuántas fueron erróneas. Esta matriz incluye:

- Verdaderos Positivos (TP). Comentarios violentos correctamente identificados.

- Falsos Positivos (FP). Comentarios no violentos incorrectamente clasificados como violentos.
- Falsos Negativos (FN). Comentarios violentos que el modelo no detectó correctamente.
- Verdaderos Negativos (TN). Comentarios no violentos correctamente identificados.

El análisis de estos resultados permitió realizar ajustes y mejoras en el modelo, optimizando su desempeño antes de su implementación en entornos reales.

5.1.3 Evaluación del rendimiento del modelo

En esta sección se evalúa el rendimiento del modelo *DistilBERT* para la clasificación de comentarios violentos y no violentos, empleando tanto métricas de pérdida durante el entrenamiento como métricas de clasificación en la evaluación. Se describen los avances logrados en las 100 épocas de entrenamiento, destacando la disminución progresiva de la pérdida como indicador de aprendizaje. Asimismo, se presentan los resultados obtenidos en el conjunto de prueba, incluyendo precisión, exhaustividad, F1-Score, exactitud y el índice Kappa de Cohen. Finalmente, se examinan los errores cometidos por el modelo, analizando los falsos positivos y falsos negativos, y se identifican áreas clave donde se puede mejorar el desempeño, como el ajuste de hiperparámetros y el aumento de datos de entrenamiento.

5.1.3.1 Rendimiento del Modelo

Para el entrenamiento, el modelo *DistilBERT* fue ajustado para clasificar comentarios como de "violencia" o no. Se entrenó en varias oportunidades, con diferentes épocas debido a las restricciones de recursos computacionales; sin embargo, en el entrenamiento final, durante 100 épocas, se observó cómo evolucionaba la pérdida (*loss*), que mide qué tan lejos estaba el modelo de la solución ideal. Al inicio, en la primera época, la pérdida fue de 0.10246395319700241, lo que mostró que el modelo aún no había aprendido

mucho. Sin embargo, en la segunda época, la pérdida bajó a 0.028653312474489212, indicando que el modelo comenzaba a mejorar y a ajustarse mejor a los datos. En la tercera época, la pérdida descendió aún más a 0.009714066982269287, lo que refleja que el modelo estaba aprendiendo de manera más eficiente y acercándose a una solución óptima. Al finalizar el entrenamiento, en la época 100 la pérdida descendió a 4.2020394175779074e-05, y se obtuvo una precisión de 0.9218, lo que significa que el modelo fue capaz de clasificar correctamente el 92.18% de los comentarios de manera correcta.

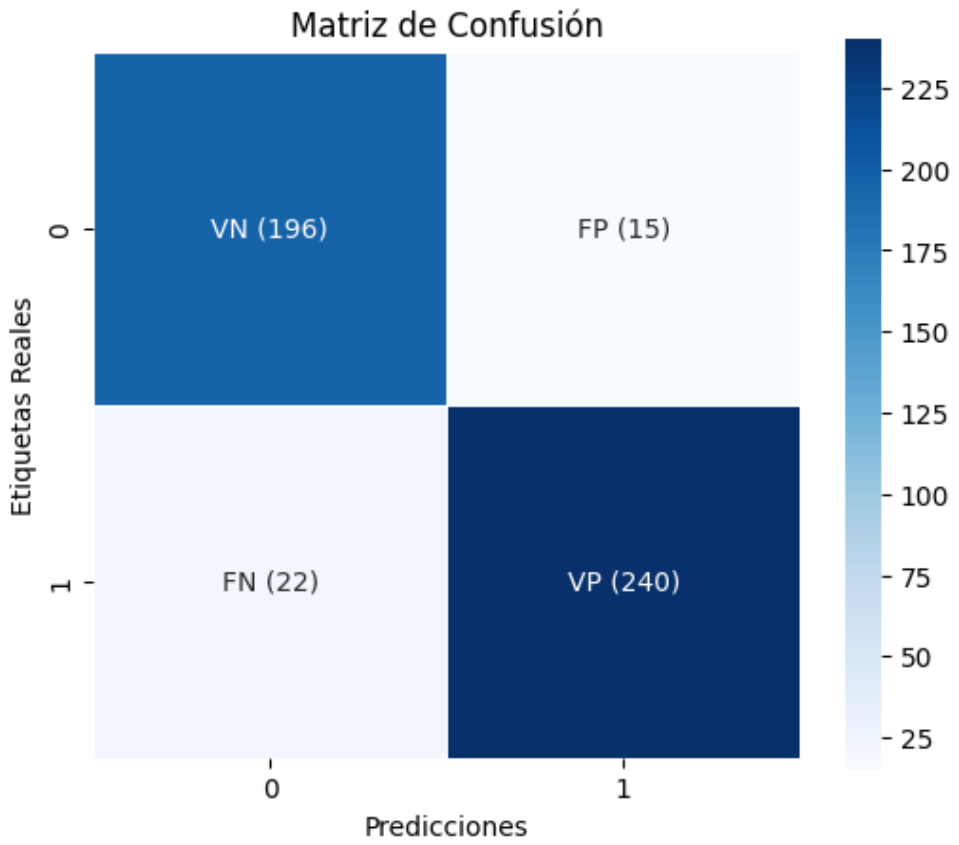
Asimismo, el modelo de clasificación fue evaluado utilizando una matriz de confusión, la cual permitió analizar el desempeño del modelo en la detección de comentarios violentos y no violentos. De acuerdo con la matriz obtenida:

- 196 comentarios no violentos fueron correctamente clasificados como no violentos (Verdaderos Negativos, VN).
- 240 comentarios violentos fueron correctamente identificados como violentos (Verdaderos Positivos, VP).
- 15 comentarios no violentos fueron erróneamente clasificados como violentos (Falsos Positivos, FP).
- 22 comentarios violentos fueron clasificados incorrectamente como no violentos (Falsos Negativos, FN).

Estos resultados (Figura 11) indican que, si bien el modelo tiene un buen desempeño en la clasificación de comentarios violentos y no violentos, aún presenta errores en la identificación de algunos comentarios violentos (FN) y ciertos casos en los que comentarios no violentos son mal clasificados como violentos (FP).

Figura 10

Matriz de confusión para la evaluación del modelo



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

A partir de los resultados obtenidos en la matriz de confusión, se calcularon las siguientes métricas de clasificación (Tabla 9):

a. **Precisión.** La precisión, que mide el porcentaje de predicciones positivas correctas, alcanzó un valor de 94.12%. Esto indica que la mayoría de los comentarios clasificados como violentos realmente contenían señales de violencia, reflejando la capacidad del modelo para minimizar falsos positivos.

b. **Exhaustividad *(Recall o Sensibilidad).*** La exhaustividad, que evalúa qué proporción de los comentarios violentos reales fueron correctamente identificados, obtuvo un valor de 91.57%. Esto sugiere que el modelo logró detectar una cantidad significativa de comentarios violentos dentro del

conjunto de prueba, aunque aún existen casos en los que no se reconoció correctamente el contenido agresivo.

c. **F1-Score.** El F1-Score, que combina precisión y exhaustividad en una sola métrica, se situó en 0.928. Este resultado confirma que el modelo mantiene un equilibrio adecuado entre la identificación de comentarios violentos y la reducción de falsos positivos, lo que contribuye a un rendimiento sólido y confiable.

d. **Exactitud *(Accuracy).*** La exactitud, que representa el porcentaje total de predicciones correctas, alcanzó un 94.05%. Este resultado demuestra que el modelo es capaz de clasificar correctamente la gran mayoría de los comentarios evaluados, asegurando un desempeño general robusto.

e. **Índice de Kappa de *Cohen.*** El índice Kappa, que mide el grado de acuerdo entre las predicciones del modelo y las etiquetas reales ajustado por el azar, obtuvo un valor de 0.88. Este resultado indica que existe un acuerdo sustancial entre el modelo y los datos reales, lo que respalda la fiabilidad de las predicciones realizadas.

Finalmente, en esta parte del estudio haciendo un análisis de errores se puede observar que, los falsos positivos (15 casos) corresponden a comentarios que, a pesar de no ser violentos, fueron clasificados erróneamente como tales. Esto podría explicarse por la presencia de términos ambiguos o contextos en los que ciertas palabras no violentas fueron interpretadas incorrectamente como señales de agresión.

Por otro lado, los falsos negativos (22 casos) representan comentarios violentos que el modelo no logró identificar correctamente. La existencia de estos errores sugiere que aún hay margen de mejora en el desempeño del modelo, ya sea mediante ajustes en los hiperparámetros, la ampliación del conjunto de datos de entrenamiento o el uso de técnicas avanzadas de procesamiento de lenguaje natural.

Tabla 9

Métricas de clasificación

Métrica	Valor	Interpretación
Precisión (Precisión)	94.12%	Mide el porcentaje de predicciones positivas correctas, reflejando la capacidad del modelo para minimizar falsos positivos.
Exhaustividad (Recall)	91.57%	Evalúa qué proporción de los comentarios violentos reales fueron correctamente identificados.
F1-Score	0.928	Combina precisión y exhaustividad en una sola métrica, mostrando el equilibrio del modelo.
Exactitud (Accuracy)	94.05%	Representa el porcentaje total de predicciones correctas, reflejando el desempeño general del modelo.
Índice de Kappa de Cohen	0.88	Mide el grado de acuerdo entre las predicciones del modelo y las etiquetas reales ajustado por el azar.

Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

5.1.3.2 Visualización del Rendimiento

La Figura 12, presentada corresponde a la Curva [ROC] *(Receiver Operating Characteristic) d*el modelo de clasificación utilizado para la detección de comentarios violentos y no violentos. Esta curva es una herramienta fundamental en la evaluación del rendimiento de modelos de clasificación binaria, ya que permite analizar su capacidad para distinguir entre ambas categorías.

En la gráfica, el eje horizontal representa la tasa de falsos positivos [FPR] (False Positive Rate), es decir, el porcentaje de comentarios no violentos que el modelo clasificó erróneamente como violentos. Por otro lado, el eje vertical muestra la tasa de verdaderos positivos [TPR] (True Positive Rate), que indica la proporción de comentarios violentos correctamente identificados.

Uno de los aspectos más relevantes de la curva ROC es su Área Bajo la Curva [AUC], (Area Under the Curve), que en este caso alcanza un valor de 0.98. Este resultado es altamente positivo, ya que un valor de 1.0 representa una clasificación perfecta, mientras que un valor de 0.5 indicaría que el modelo no tiene

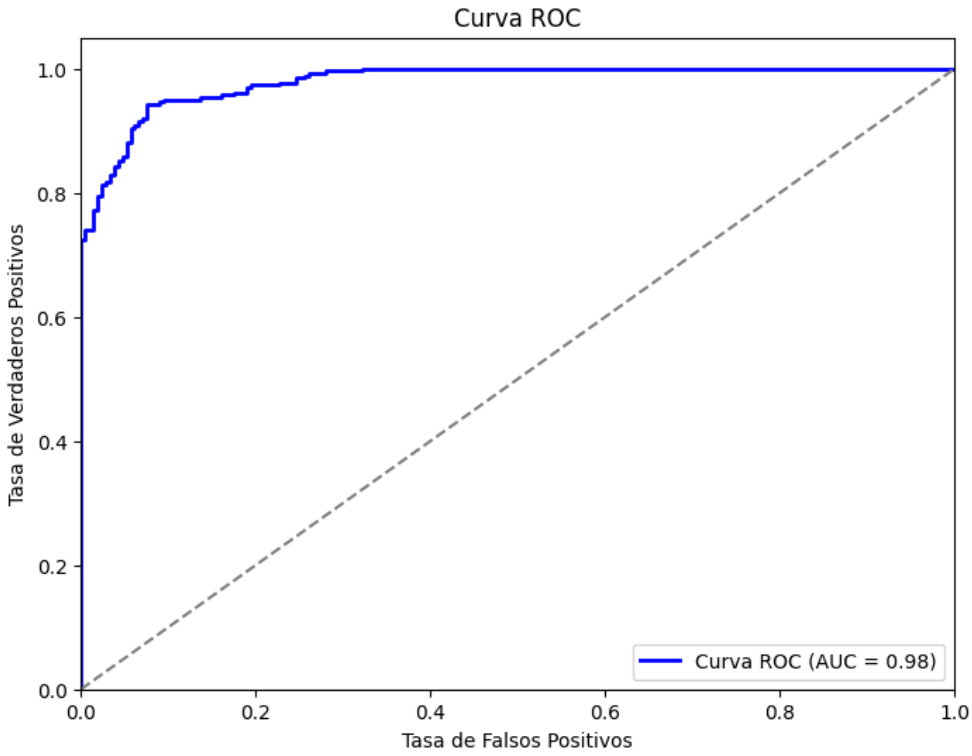
capacidad de diferenciación, clasificando los datos de manera aleatoria. El hecho de que el modelo evaluado obtenga un AUC de 0.98 sugiere que posee una capacidad excepcional para identificar correctamente los comentarios violentos sin etiquetar erróneamente demasiados comentarios no violentos.

Además, la curva se encuentra muy por encima de la diagonal punteada, lo que indica que el modelo logra una alta tasa de detección de comentarios violentos (sensibilidad) al tiempo que minimiza los falsos positivos. En otras palabras, tiene un rendimiento sobresaliente en la clasificación, diferenciando de manera precisa entre mensajes con contenido violento y aquellos que no presentan indicios de agresión verbal.

Este resultado demuestra que el modelo es una herramienta eficaz para la detección automatizada de señales de violencia en redes sociales, proporcionando predicciones confiables en la mayoría de los casos. No obstante, si se requiere una optimización adicional, se pueden realizar ajustes en el umbral de clasificación para equilibrar mejor la detección de verdaderos positivos y la reducción de falsos positivos, dependiendo de las necesidades específicas del análisis.

Figura 11

Visualización del rendimiento a través de la curva ROC



Fuente: Bastidas, Crose & Lopez (2024). Detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

5.1.4 Herramienta digital para el análisis de patrones de violencia

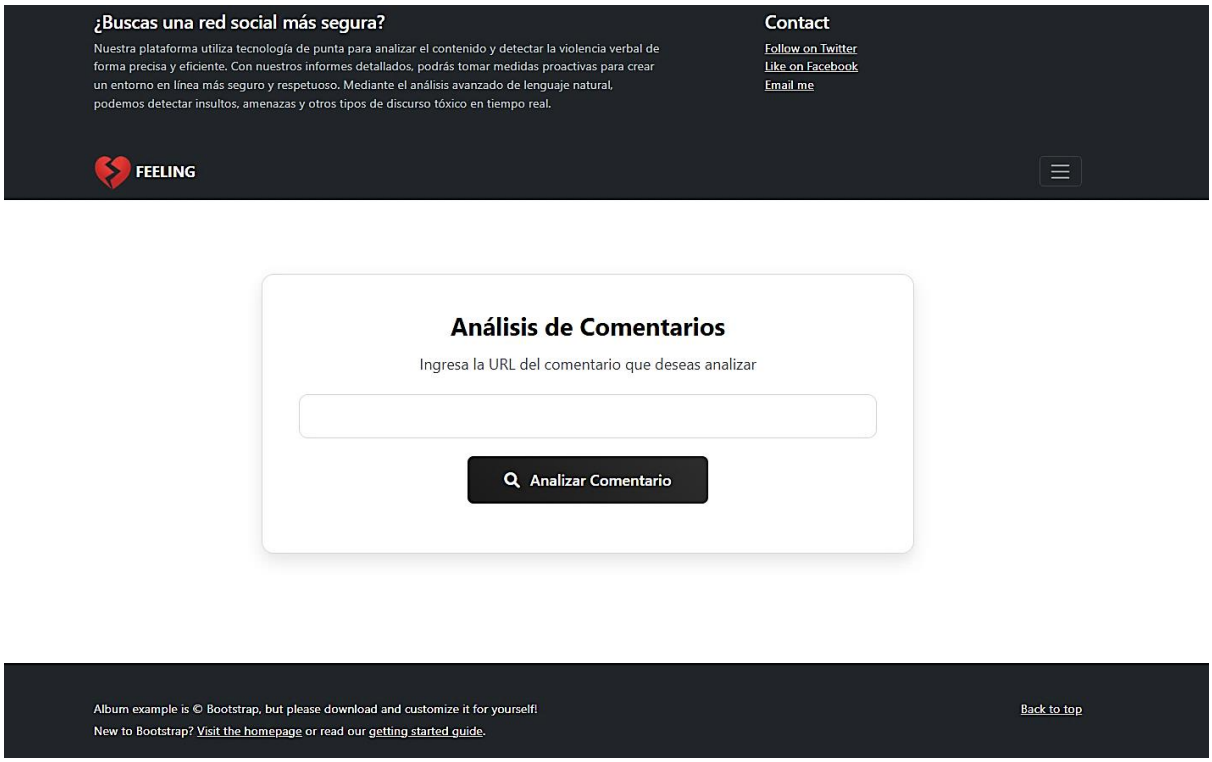
En esta sección se presenta una herramienta digital diseñada para el análisis de patrones de violencia verbal en comentarios de redes sociales, específicamente en *YouTube*. Esta aplicación ha sido desarrollada utilizando el *framework Django*, mientras que la visualización de datos se lleva a cabo mediante *Plotly*, lo que permite una representación gráfica interactiva y dinámica de los resultados.

5.1.4.1 Descripción de la funcionalidad de la herramienta.

Con el objetivo de facilitar la detección y análisis de patrones de violencia verbal en redes sociales, se desarrolló una herramienta digital basada en técnicas de procesamiento de lenguaje natural y aprendizaje automático. Esta plataforma permite a los usuarios ingresar enlaces de publicaciones en *YouTube* (Figura 13), extrayendo comentarios y clasificándolos automáticamente como violentos o no violentos.

Figura 12

Plataforma para la introducción de enlaces de You Tube

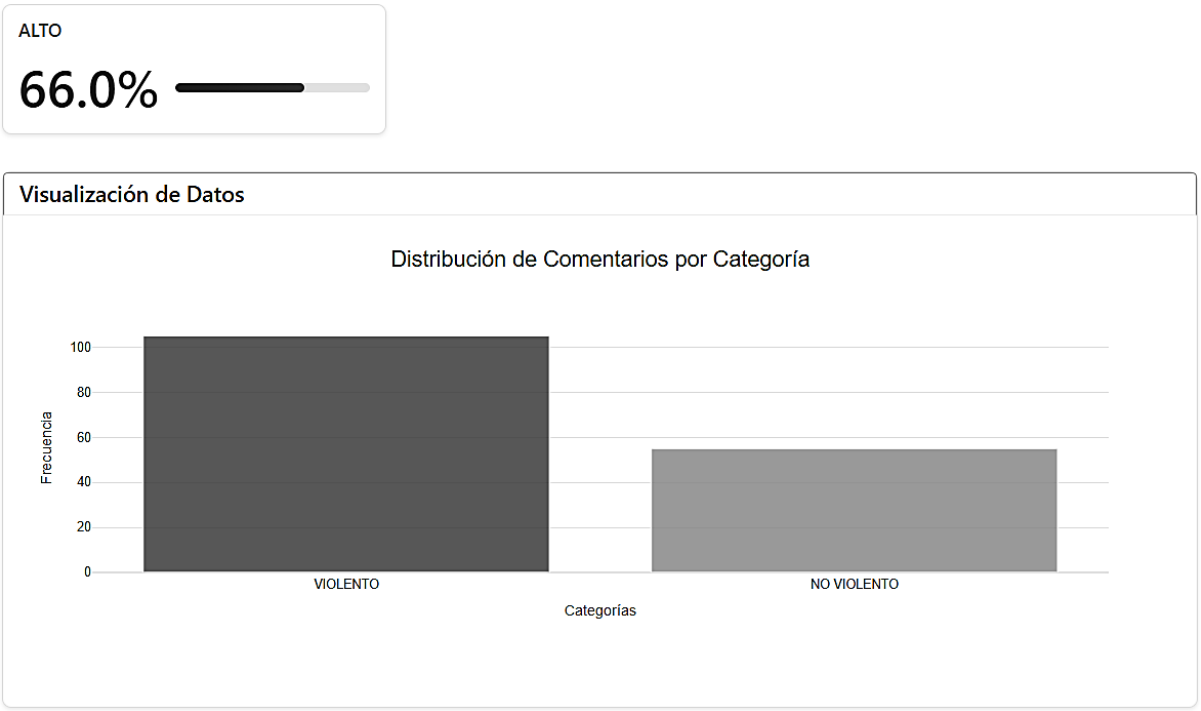


Fuente: Bastidas, Crose & Lopez (2024). Plataforma para la introducción de enlaces para la detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

A través de gráficos interactivos (Figura 14), y reportes detallados, la herramienta brinda una representación visual de la distribución de comentarios, proporcionando información clave sobre la prevalencia de lenguaje agresivo en diversos contextos. Estos términos se organizan de acuerdo con su frecuencia de aparición, permitiendo una comprensión clara de los patrones lingüísticos predominantes en los discursos agresivos.

Figura 13

Interactivos que resaltan las palabras más violentas usadas en los comentarios



Fuente: Bastidas, Crose & Lopez (2024). Plataforma para la introducción de enlaces para la detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

Para optimizar la visualización de los datos, la aplicación incorpora gráficos interactivos, como gráficos de barras y paneles dinámicos, que facilitan la interpretación de los resultados. En el caso de grandes volúmenes de comentarios, el sistema selecciona aleatoriamente 10 comentarios representativos, mientras que, si el número de comentarios es menor a 10, se presentan todos los disponibles (Figura 15). Además, se muestra el número de palabras violentas por comentario, así como la proporción de comentarios clasificados como violentos o no violentos, expresada en términos porcentuales.

Figura 14

Métricas evaluadas en la plataforma web.

Datos Detallados

comentario_limpio	numero_palabras	ocurrencia	predicciones
Dina pinocha como te quedo la nariz operada	8	1	0
Y sin papel y no como la burra de dina	10	1	1
Dinauna más de sus farsas	5	0	1
Cuando no la prensa de obredech distorcionando	7	0	0
Siga adelante Sra Boluarte Ya no quiero contrariarla Sería bueno que traiga productos chinos de alta gama al Perú a buen precio Eso va a cambiar la percepción del pueblo Y pida consejos con mucha humildad a los presidentes que Ud conoce Consejo de conejo	45	0	1
Por favor dejense de tanto odió y resentimiento dedicarse a trabajar	11	0	0
Los encargados de construir penales NO ES LA SOCIEDAD CIVIL EN NINGUN PAIS DELMUNDO ES ASI	16	1	1
Estos rojos nunca sabrán valorar todo proyecto es largo los que comienzan no inaugura sino el que lo sucede aprendan a dar el valor correspondiente si no lo sigue ahí se pierde lo que vale es haber seguido hasta su culminación eso hay que resaltar	45	0	1
FUERA MTODO ES ARREGLADO CUÁNTO LIMOSNA HABRÁ RECIBIDO LA DE BLUSITA VERDE	12	0	1
SI ES DEL PUEBLO POR QUE NO LO HICIERON ANTES LOS CASTIBURROS Y SUS BESTIAS ESTÁN FURIOSOS	17	3	0

Anterior 1 2 3 4 5 Siguiente

Fuente: Bastidas, Crose & Lopez (2024). Plataforma para la introducción de enlaces para la detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

Esta solución digital combina eficiencia y facilidad de uso, proporcionando una plataforma robusta para la detección y el análisis de violencia verbal en redes sociales. Su diseño intuitivo y la integración de herramientas analíticas avanzadas permiten a los usuarios explorar tendencias y patrones en el lenguaje empleado en entornos digitales, contribuyendo al estudio y mitigación del discurso agresivo en línea.

5.1.4.2 Proceso de Implementación, funcionamiento y evaluación de desempeño.

Para la recolección de datos, se implementó un módulo de Web Scraping que automatiza la extracción de comentarios desde publicaciones de YouTube. Este proceso permite capturar texto sin necesidad de intervención manual, almacenando los comentarios en una base de datos estructurada. A partir de este

proceso, los datos son limpiados para eliminar caracteres especiales, emoticones y estructuras no relevantes para el análisis de lenguaje (Figura 16).

Figura 15

Automatización de Extracción de Comentarios en YouTube mediante Web Scraping con Selenium y Pandas

```
from selenium import webdriver
from selenium.webdriver.common.by import By
import pandas as pd
from selenium.webdriver.firefox.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.firefox.options import Options
from webdriver_manager.firefox import GeckoDriverManager
import time


options = Options()
options.headless = True
service = Service(GeckoDriverManager().install())
driver = webdriver.Firefox(service=service, options=options)


def youtube(urls):
    comentarios = []  # Initialize an empty list to store comments

    # Assuming `driver` is already initialized before this function is called

    for url in urls:
        driver.get(url)
        time.sleep(5)

        # Scroll and collect comments
        for _ in range(16):
            try:
                WebDriverWait(driver, 5).until(
                    EC.visibility_of_element_located((By.ID, 'comments'))
                )
                time.sleep(3)
                comments = driver.find_elements(By.ID, 'comments')

                if comments:
                    driver.execute_script("window.scrollBy(0, 800);")
            except Exception as e:
                print(f"Error: {e}")
                break

        # Extract the comments
        elementos = driver.find_elements(By.CSS_SELECTOR, 'ytd-comment-thread-renderer')

        for elemento in elementos:
            try:
                comentario_text = elemento.find_element(By.CSS_SELECTOR, '#content-
text').text
                comentarios.append(comentario_text)  # Add the comment to the list
                print("->: ",comentario_text)
            except Exception as e:
                print(f"No se pudo encontrar el comentario: {e}")

    # Create a DataFrame from the collected comments
    df = pd.DataFrame(comentarios, columns=["Comentario"])

    # Remove duplicates based on the comment text
    dataset = df.drop_duplicates(subset="Comentario")

    # Write the dataset to a CSV file
    dataset.to_csv('Youtubedataset.csv', index=False, encoding='utf-8' , sep = ',')
```

Fuente: Bastidas, Crose & Lopez (2024). Plataforma para la introducción de enlaces para la detección de señales de violencia en redes sociales: un enfoque desde la ciencia de datos.

Cada comentario extraído es transformado en una representación numérica mediante un proceso de tokenización basado en DistilBERT. Posteriormente, el modelo clasifica los textos en dos categorías: comentarios con contenido violento y comentarios sin indicios de agresión verbal. Para este proceso, se asigna una etiqueta binaria a cada observación, permitiendo la identificación automática de patrones de lenguaje agresivo en los datos analizados.

Una vez finalizado el análisis, la herramienta presenta los resultados mediante gráficos interactivos y reportes tabulares. El gráfico de barras muestra la proporción de comentarios violentos y no violentos, permitiendo a los usuarios observar la distribución de la violencia verbal en los datos analizados. Además, se presenta una tabla detallada con los comentarios procesados, el número de palabras en cada uno, su clasificación y ocurrencia. Esta visualización facilita la interpretación de los resultados y permite detectar patrones específicos de lenguaje agresivo.

5.1.4.3 Análisis de Resultados y Evaluación del Desempeño

El análisis de desempeño del modelo sugiere que la herramienta alcanza una precisión del 94.12% en la clasificación de comentarios violentos, con una tasa de falsos positivos del 5.88%. Aunque el modelo logra identificar correctamente la mayoría de los comentarios agresivos, se observan algunos errores en textos cuyo significado depende del contexto. Estos hallazgos destacan la necesidad de continuar optimizando el modelo, incorporando estrategias que permitan mejorar la interpretación semántica del contenido analizado

5.1.4.4 Limitaciones y Posibles Mejoras

Si bien la herramienta desarrollada ha demostrado ser efectiva en la detección de violencia verbal en redes sociales, presenta ciertas limitaciones relacionadas con el análisis semántico del contexto. Comentarios irónicos o ambiguos pueden ser clasificados erróneamente, lo que sugiere la necesidad de integrar técnicas avanzadas de análisis de contexto. Asimismo, el modelo podría beneficiarse de un conjunto de datos más amplio que capture una mayor diversidad lingüística y cultural. Futuras investigaciones podrían extender la funcionalidad de

la herramienta a otras plataformas, ampliando así su aplicabilidad en diferentes entornos digitales. Por ejemplo:

- Contexto semántico: Algunas expresiones pueden ser malinterpretadas como violentas debido a la falta de comprensión del tono o sarcasmo en los comentarios.

- Dependencia del dataset de entrenamiento: La herramienta podría mejorar su rendimiento al entrenarse con una base de datos más amplia y diversa.

- Extensibilidad a otras plataformas: Actualmente, la herramienta se centra en YouTube, pero podría adaptarse a otras redes sociales como X o Instagram.

5.2 Discusión de resultados

En el presente estudio, se planteó como primer objetivo específico 1, Identificar y analizar las características y patrones de lenguaje en comentarios de redes sociales que contienen violencia verbal, mediante la extracción, etiquetado y preprocesamiento de un conjunto de datos. En tal sentido, la recolección de datos se llevó a cabo mediante la técnica de Web Scraping, extrayendo un total de 9556 comentarios de Facebook y YouTube, provenientes de 1646 enlaces seleccionados aleatoriamente. Estos datos abarcaban una amplia gama de contextos, incluyendo política, deporte, economía, psicología, problemáticas sociales y entretenimiento. Para garantizar la calidad del conjunto de datos, se implementó un primer proceso de filtrado, eliminando comentarios con emoticones, textos excesivamente cortos o caracteres atípicos, lo que redujo la base a 4507 comentarios.

Posteriormente, se llevó a cabo una clasificación semántica supervisada por un lingüista, quien identificó y eliminó comentarios ambiguos, es decir, aquellos con términos groseros cuyo significado variaba según el contexto. Como resultado de este proceso, la muestra final quedó compuesta por 2343 comentarios, clasificados binariamente como violentos (1) o no violentos (0).

El estudio reveló que la mayoría de los comentarios en redes sociales son breves y concisos. Se observó que el 89.54% de los comentarios contienen entre 1 y 33 palabras, lo que sugiere una fuerte tendencia hacia mensajes cortos y

directos. En contraste, los comentarios de mayor extensión (más de 65 palabras) representaron apenas el 1.49% del total.

En términos de medidas de resumen, los 2343 comentarios analizados presentan una longitud media de 20.79 palabras, con una desviación estándar de 22.94, lo que indica una variabilidad considerable en la extensión de los textos. La distribución es altamente asimétrica, con una curtosis de 113.47 y un sesgo de 8.98, lo que significa que la mayoría de los comentarios son cortos, pero existen algunos casos atípicos con una gran cantidad de palabras.

Estos resultados tienen implicaciones clave en la detección automatizada de violencia en redes sociales, ya que los modelos de PLN deben estar optimizados para manejar textos predominantemente breves. Esto resalta la importancia de técnicas de tokenización adecuadas y representaciones vectoriales eficientes para mejorar la clasificación de comentarios en estos entornos digitales.

El estudio también examinó la frecuencia de términos violentos en los comentarios clasificados como agresivos. Se encontró que la gran mayoría de estos comentarios (89.54%) contenían entre 1 y 33 términos violentos, lo que indica que la violencia verbal en redes sociales tiende a manifestarse en textos relativamente cortos.

La distribución de los términos violentos reveló que, aunque la media de palabras violentas por comentario fue de 1.37, la desviación estándar de 1.95 sugiere que existen casos con una cantidad significativamente mayor de términos agresivos. La presencia de valores extremos es evidente, con comentarios que contienen hasta 19 términos violentos, lo que resalta la heterogeneidad del fenómeno analizado.

Asimismo, el análisis semántico permitió la construcción de un diccionario especializado de términos violentos, basado en la frecuencia de aparición de palabras dentro de los comentarios clasificados como agresivos. Entre los términos más comunes se identificaron "corrupto", "delincuente", "criminal", "rata" y "asesino", lo que sugiere que la violencia verbal en redes sociales está fuertemente vinculada a temáticas relacionadas con la corrupción y la criminalidad.

Estos hallazgos son fundamentales para el diseño de sistemas automatizados de detección de violencia en redes sociales, ya que proporcionan una base léxica sobre la cual se pueden construir modelos de clasificación más eficientes.

Cabe resaltar que, el análisis de la distribución de los comentarios clasificados reveló una proporción notablemente equilibrada entre comentarios violentos (49.13%) y no violentos (50.87%). Esta relación difiere de estudios previos, donde los comentarios agresivos representaban una fracción significativamente menor del total.

La alta incidencia de comentarios violentos en la muestra analizada pone en evidencia la relevancia del problema y la necesidad de implementar herramientas de monitoreo y moderación automatizada para identificar discursos de odio y mitigar su impacto en los entornos digitales.

Al comparar los resultados de nuestro estudio con los hallazgos de Chatzakou et al. (2017) en "Mean Birds: Detecting Aggression and Bullying on Twitter", se observan tanto similitudes como diferencias significativas en la detección de comportamientos agresivos en redes sociales. Mientras que en el presente estudio se tiene un enfoque que se centró en la extracción y preprocesamiento de comentarios de YouTube para identificar señales de violencia verbal, Chatzakou et al. dirigieron su investigación hacia la detección de comportamientos agresivos y de acoso en Twitter.

En términos metodológicos, ambos estudios emplearon técnicas de análisis de texto y aprendizaje automático para clasificar el contenido. Sin embargo, Chatzakou et al. implementaron un enfoque más integral al considerar atributos textuales, del usuario y de la red, logrando una precisión en la detección de usuarios agresivos superior al 90%. Por otro lado, el presente estudio con una precisión de 94.12% se enfocó en el análisis léxico y la identificación de términos violentos específicos en los comentarios, lo que permitió una comprensión detallada de la narrativa violenta en las plataformas analizadas.

Una diferencia notable radica en las plataformas estudiadas. Twitter, con su límite de caracteres y naturaleza pública, facilita interacciones más inmediatas y, a menudo, más impulsivas, lo que puede influir en la prevalencia y naturaleza del contenido agresivo. En contraste, YouTube permite publicaciones más extensas y pueden estar sujetas a diferentes dinámicas de interacción, lo que podría explicar las variaciones en la proporción de contenido violento identificado.

Además, mientras que nuestro estudio encontró que aproximadamente el 49.1% de los comentarios contenían señales de violencia, Chatzakou et al. no proporcionaron un porcentaje global de contenido abusivo, sino que se enfocaron en caracterizar las propiedades de los usuarios agresivos y cómo se distinguen de los usuarios regulares.

Finalmente, aunque ambos estudios abordan la detección de comportamientos agresivos en redes sociales, difieren en sus enfoques metodológicos y en las plataformas analizadas. Estas diferencias resaltan la importancia de adaptar las estrategias de detección al contexto específico de cada plataforma y sugieren que una combinación de análisis léxico detallado y técnicas avanzadas de aprendizaje automático podría mejorar la identificación de señales de violencia verbal en diversos entornos digitales.

El objetivo específico 2 del presente estudio fue desarrollar un modelo de análisis de sentimientos para detectar violencia verbal con técnicas de PLN; es decir, este se centró en la detección de señales de violencia verbal en redes sociales mediante el análisis de sentimientos y técnicas avanzadas de PLN. Para ello, se implementó el modelo DistilBERT, una versión optimizada y eficiente de BERT, desarrollada por Sanh et al. (2019). Este modelo conserva el 97% del rendimiento de BERT, pero con una reducción del 40% en su tamaño, lo que lo hace más rápido y eficiente en tareas de clasificación de texto. Su capacidad para analizar el contexto bidireccional de las palabras y capturar relaciones semánticas lo convierte en una herramienta ideal para la detección automatizada de discursos agresivos en plataformas digitales.

El proceso de implementación del modelo inició con la recopilación de datos a partir de comentarios extraídos de Facebook y YouTube, utilizando como ya se

mencionó la técnica de Web Scraping. Para garantizar la precisión en la clasificación, se realizó un análisis semántico supervisado por un lingüista, lo que permitió la depuración final de 2343 comentarios, clasificados como violentos (1) o no violentos (0).

Para la implementación del modelo DistilBERT, se estableció un procedimiento estructurado que abarcó varias etapas clave. En primer lugar, se organizó la base de datos en una estructura matricial, facilitando la trazabilidad de cada comentario mediante un identificador único y etiquetas de clasificación. Posteriormente, se aplicó un preprocesamiento lingüístico que incluyó la eliminación de caracteres especiales y la conversión de los textos en representaciones numéricas mediante técnicas como TF-IDF, Word2Vec y BERT, lo que permitió al modelo comprender mejor el significado de los textos.

Una vez procesados los datos, se dividieron en tres conjuntos: entrenamiento (80%), validación (10%) y prueba (10%), asegurando un adecuado balance para la evaluación del modelo. Durante la fase de entrenamiento, DistilBERT fue ajustado mediante fine-tuning, optimizando hiperparámetros como la tasa de aprendizaje, el número de épocas y el tamaño del lote. Este ajuste permitió mejorar la capacidad del modelo para identificar patrones lingüísticos asociados a la violencia verbal.

En decir, la presente investigación valida la eficacia de DistilBERT en la detección automatizada de discursos agresivos en redes sociales, demostrando que el modelo puede diferenciar con alta precisión entre lenguaje violento y no violento. La combinación de técnicas avanzadas de preprocesamiento, tokenización y representación numérica del lenguaje optimizó su rendimiento y mejoró su aplicabilidad en entornos de monitoreo de contenido digital. En ese sentido, estos hallazgos se pueden comparar con el estudio de Atapattu et al. (2020), quienes aplicaron modelos de PLN para la detección de ciberacoso y discurso de odio en la plataforma Twitter.

Uno de los aspectos más relevantes en esta comparación es el rendimiento del modelo en términos de precisión y recall. En el presente estudio, se obtuvo una precisión del 94.12%, lo que indica que casi la totalidad de los comentarios

clasificados como violentos realmente contenían señales de agresión verbal. En contraste, el modelo de Atapattu et al. alcanzó una precisión del 73% en la detección de discursos de odio y 74% en la identificación de agresividad. Esta diferencia sugiere que el modelo basado en DistilBERT, empleado en este estudio, presenta una mayor capacidad para minimizar falsos positivos, lo que se traduce en una clasificación más confiable.

Asimismo, la exhaustividad o recall en este estudio alcanzó un 91.57%, reflejando la capacidad del modelo para detectar la mayoría de los comentarios violentos presentes en el conjunto de datos. En el caso de Atapattu et al., esta métrica fue significativamente menor, lo que indica que su modelo tuvo más dificultades para identificar algunos casos de violencia verbal. Esta diferencia podría deberse a la estructura del lenguaje en las plataformas analizadas: mientras que Twitter se caracteriza por mensajes cortos y con menos contexto, en YouTube los comentarios suelen ser más largos, lo que facilita la identificación de patrones semánticos y lingüísticos que distinguen el lenguaje violento del no violento.

En términos de exactitud general (accuracy), el presente estudio obtuvo un 94.05%, lo que indica que el modelo logró clasificar correctamente una gran mayoría de los comentarios analizados. En el caso del estudio de Atapattu et al., no se proporciona un valor exacto de esta métrica, pero los resultados sugieren que su modelo presentó mayores dificultades en la clasificación correcta de ciertos comentarios, lo que podría deberse a una menor calidad del preprocesamiento de datos o a la variabilidad en los discursos analizados.

Además, se evaluó la concordancia entre las predicciones del modelo y las etiquetas reales mediante el índice de Kappa de Cohen, obteniéndose un valor de 0.88. Este resultado sugiere que la clasificación del modelo tiene un alto nivel de consistencia y fiabilidad, minimizando la influencia del azar en las decisiones del sistema. En comparación, aunque Atapattu et al. no reportaron esta métrica en sus hallazgos, mencionan que su modelo presentó dificultades en la clasificación de ciertos discursos ambiguos, lo que sugiere una menor concordancia entre sus predicciones y los datos reales.

Otra diferencia clave entre ambos estudios radica en el preprocesamiento de datos y la clasificación de comentarios. En este estudio, se realizó un análisis semántico supervisado por un lingüista, permitiendo una depuración más precisa de los datos y la eliminación de comentarios ambiguos, lo que contribuyó a mejorar la calidad del conjunto de entrenamiento del modelo. En contraste, Atapattu et al. basaron su clasificación en un enfoque puramente automatizado, lo que pudo haber introducido mayor ruido en los datos y afectado la precisión de su modelo.

No obstante, una de las principales fortalezas del trabajo de Atapattu et al. es que evaluaron la adaptabilidad de su modelo en diferentes conjuntos de datos, obteniendo un desempeño estable con un F1-Score de 0.7 en dominios externos. En cambio, el presente estudio no realizó pruebas de generalización fuera del conjunto de datos original, lo que representa una oportunidad de mejora para futuras investigaciones. Evaluar el modelo en distintos entornos permitiría determinar si su alto desempeño se mantiene en otras plataformas o comunidades digitales.

Finalmente, el modelo desarrollado en este estudio supera al de Atapattu et al. en términos de precisión (94.12% vs. 73%), recall (91.57% vs. valores más bajos no especificados) y F1-Score (0.928 vs. 0.74), lo que indica un mayor equilibrio entre la detección de comentarios violentos y la reducción de falsos positivos. Asimismo, la alta exactitud (94.05%) y el índice de Kappa (0.88) reflejan que el modelo es altamente confiable y presenta una mínima variabilidad en sus predicciones. Sin embargo, una limitación del presente estudio es la falta de validación del modelo en datos externos, mientras que Atapattu et al. exploraron su adaptabilidad a distintos dominios.

Por lo tanto, aunque el presente modelo basado en DistilBERT demuestra una capacidad superior para la detección de violencia verbal en redes sociales, futuros estudios podrían expandir su evaluación a nuevas plataformas y explorar técnicas adicionales de aprendizaje transferido, con el fin de mejorar su aplicabilidad en diferentes contextos lingüísticos y sociales.

Por otro lado, cabe resaltar la comparación del presente estudio con el de, Campillo Muñoz (2022), quién adoptó un enfoque basado en lingüística de corpus

para analizar la violencia verbal en redes sociales, con especial énfasis en el análisis de pragmática y estructura discursiva. Aunque este trabajo no se centró exclusivamente en modelos de PLN como DistilBERT, sus hallazgos sobre la estructura y patrones del lenguaje violento pueden complementar los resultados del presente estudio.

Una coincidencia importante entre ambos estudios es la identificación de patrones lingüísticos comunes en discursos agresivos, como el uso de insultos directos, términos despectivos y estructuras sintácticas específicas. En el presente trabajo, se emplearon técnicas de tokenización y representación numérica del lenguaje, como TF-IDF, Word2Vec y BERT, para capturar estas características en comentarios de redes sociales. Campillo Muñoz, por otro lado, realizó un análisis cualitativo y cuantitativo de términos violentos en interacciones digitales, encontrando que la violencia verbal en línea a menudo se enmarca en contextos de descalificación política, ataques personales y discursos polarizantes.

En cuanto a la metodología, el presente estudio adoptó un enfoque computacional automatizado, mientras que Campillo Muñoz empleó métodos manuales y estadísticos para el análisis de corpus lingüísticos. A pesar de estas diferencias, ambos estudios coinciden en que la violencia verbal en redes sociales sigue estructuras lingüísticas predecibles, lo que refuerza la importancia de integrar análisis de corpus en modelos de PLN para mejorar la clasificación de contenido agresivo.

Una diferencia clave entre ambos estudios es que Campillo Muñoz no aborda la clasificación automática de textos, sino que se enfoca en la comprensión del fenómeno lingüístico, mientras que en el presente trabajo se implementó un modelo de aprendizaje profundo basado en redes neuronales. Sin embargo, los hallazgos de Campillo Muñoz podrían complementar el proceso de preprocesamiento del lenguaje, ayudando a mejorar la identificación de términos violentos y su contexto semántico dentro de modelos como DistilBERT.

El objetivo específico 3 del presente estudio es evaluar y optimizar el rendimiento del modelo con métricas de clasificación; es decir, se evaluó el rendimiento del modelo DistilBERT para la clasificación de comentarios violentos y

no violentos en redes sociales, empleando métricas de aprendizaje y clasificación. A lo largo del proceso de entrenamiento, que abarcó 100 épocas, se observó una reducción progresiva de la pérdida del modelo (loss), indicando una mejora en su capacidad de generalización. Desde una pérdida inicial de 0.1024 en la primera iteración, el modelo logró optimizar su aprendizaje hasta alcanzar un valor de 4.2e-05 en la última época, reflejando un ajuste preciso a los datos analizados.

Para evaluar el desempeño del modelo en el conjunto de prueba, se emplearon métricas de clasificación estándar, obteniendo resultados altamente favorables. En términos de precisión, el modelo alcanzó un 94.12%, indicando que la gran mayoría de los comentarios clasificados como violentos contenían efectivamente señales de agresión. La exhaustividad (recall) fue del 91.57%, lo que demuestra que el sistema identificó correctamente la mayoría de los comentarios violentos, aunque aún existen casos que no fueron reconocidos. El F1-Score, que equilibra la precisión y la exhaustividad, se situó en 0.928, reflejando un rendimiento robusto en la clasificación binaria.

Además, el modelo obtuvo una exactitud global (accuracy) del 94.05%, indicando que logró clasificar correctamente la gran mayoría de los comentarios en el conjunto de prueba. Para evaluar la confiabilidad de las predicciones, se calculó el índice de Kappa de Cohen, alcanzando un 0.88, lo que representa un alto grado de concordancia entre las etiquetas reales y las predicciones del modelo, ajustado por el azar.

Un análisis más detallado a través de la matriz de confusión permitió identificar áreas de mejora. Se observó que 196 comentarios no violentos fueron clasificados correctamente (Verdaderos Negativos, VN) y 240 comentarios violentos fueron identificados con precisión (Verdaderos Positivos, VP). Sin embargo, se registraron 15 falsos positivos (FP), donde comentarios no violentos fueron etiquetados erróneamente como violentos, posiblemente debido a la presencia de términos ambiguos o en contextos particulares. Asimismo, se detectaron 22 falsos negativos (FN), lo que indica que algunos comentarios violentos no fueron correctamente clasificados, sugiriendo que el modelo podría

beneficiarse de ajustes en los hiperparámetros o la ampliación del conjunto de entrenamiento.

Para analizar el rendimiento del modelo desde un enfoque visual, se generó la Curva ROC (Receiver Operating Characteristic), la cual mostró una Área Bajo la Curva (AUC) de 0.98. Este resultado es altamente positivo, ya que indica que el modelo posee una capacidad excepcional para diferenciar entre comentarios violentos y no violentos. La gráfica mostró una curva situada muy por encima de la diagonal de referencia, lo que confirma que el modelo logra una alta tasa de detección de violencia minimizando los falsos positivos.

En términos de optimización, los resultados sugieren que el modelo ya presenta un rendimiento destacado, pero aún existen oportunidades de mejora. Algunas estrategias incluyen el ajuste del umbral de clasificación, la expansión del conjunto de datos de entrenamiento con ejemplos más diversos, y la implementación de técnicas avanzadas de procesamiento de lenguaje natural, como modelos con atención contextual más refinada.

Luego, los hallazgos indican que DistilBERT es una herramienta eficaz para la detección automatizada de violencia verbal en redes sociales, con un desempeño sólido y confiable. Sin embargo, el análisis de errores sugiere que futuras optimizaciones podrían mejorar la clasificación de casos ambiguos y reducir la tasa de falsos negativos, lo que permitiría una aplicación más precisa y generalizable en distintos entornos digitales.

En seguida, para establecer un marco comparativo, se ha seleccionado el estudio de Carvajal (2023), quién reportó una precisión media de aproximadamente 92.5%, con variaciones según el modelo utilizado y el tipo de texto analizado. En particular, el modelo DistilBERT mostró una precisión cercana al 93%, lo que lo posiciona de manera comparable con los resultados obtenidos en este estudio para la detección de violencia verbal en redes sociales. Sin embargo, una diferencia clave radica en la distribución de los datos y el tipo de clasificación. Mientras que el presente estudio se centró en una tarea de clasificación binaria (violento vs. no violento), el trabajo de Carvajal abordó una clasificación multiclase, identificando distintos tipos de mensajes asociados a los ODS.

Otro aspecto a destacar es la estrategia de preprocesamiento y optimización. En este estudio, el preprocesamiento de los datos incluyó técnicas como eliminación de caracteres especiales, tokenización y representación numérica mediante TF-IDF y Word2Vec, mientras que Carvajal (2023) utilizó enfoques más avanzados de normalización semántica y embeddings contextualizados para mejorar la calidad del texto antes del entrenamiento del modelo. Esto sugiere que, si bien los dos estudios implementaron modelos de PLN de última generación, sus diferencias en la estructura de los datos y los enfoques de preprocesamiento pudieron influir en la variabilidad de sus métricas de rendimiento.

Finalmente, en términos de validación del modelo, una limitación del presente estudio es la falta de evaluación en datos externos, lo que restringe su aplicabilidad a otros contextos o plataformas digitales. En contraste, el estudio de Carvajal (2023) validó sus modelos en diferentes corpus de texto, asegurando su adaptabilidad a distintos dominios temáticos. Esto indica que una posible mejora en futuros trabajos sería evaluar el modelo en conjuntos de datos más diversos, permitiendo comprobar su capacidad de generalización en otros espacios digitales más allá de YouTube.

Luego se puede resaltar que, ambos estudios demuestran la eficacia de los modelos de Deep Learning en la clasificación de textos, con desempeños comparables en términos de precisión y robustez del modelo. Sin embargo, mientras que el presente estudio se enfoca en la detección de violencia verbal en redes sociales, el trabajo de Carvajal (2023) abarca una problemática más amplia dentro del análisis de textos en el contexto de los Objetivos de Desarrollo Sostenible. La comparación de estos enfoques resalta la versatilidad de los modelos basados en transformadores y la importancia de adaptar las metodologías según la naturaleza del problema a resolver.

En cuanto al objetivo específico 4 del presente estudio es diseñar una herramienta digital que genere informes y ventanas emergentes que muestren patrones de violencia verbal; es por ello, que se desarrolló una herramienta digital con el propósito de identificar y analizar patrones de violencia verbal en

comentarios de redes sociales, específicamente en YouTube. Para su implementación, se utilizó el framework Django como base de desarrollo y Plotly para la visualización interactiva de los datos. Esta plataforma permite a los usuarios ingresar enlaces de publicaciones en YouTube, extrayendo automáticamente los comentarios y clasificándolos como violentos o no violentos mediante un modelo de PLN basado en DistilBERT.

El sistema facilita la interpretación de los resultados mediante gráficos de barras interactivos y reportes tabulares que resumen la distribución de los comentarios analizados. En el caso de grandes volúmenes de datos, la herramienta selecciona aleatoriamente 10 comentarios representativos para su visualización, mientras que, si la cantidad es menor, muestra todos los comentarios disponibles. Además, el sistema proporciona detalles sobre la frecuencia de términos violentos dentro de cada comentario y el porcentaje total de publicaciones agresivas.

La implementación de la herramienta se basó en un módulo de Web Scraping, que automatizó la extracción de comentarios desde YouTube y los almacenó en una base de datos estructurada. Posteriormente, estos datos fueron limpiados y tokenizados para su análisis, clasificándolos en función de su carga agresiva. Los resultados de clasificación se presentan en una interfaz intuitiva que permite explorar de manera eficiente los patrones de lenguaje utilizados en interacciones digitales.

En términos de desempeño, la herramienta alcanzó una precisión del 94.12%, con una tasa de falsos positivos del 5.88%. No obstante, se identificaron limitaciones en la capacidad del modelo para interpretar correctamente el contexto semántico, especialmente en casos de ironía o comentarios ambiguos. Asimismo, se observó la necesidad de expandir el conjunto de datos de entrenamiento para mejorar la diversidad lingüística del sistema y garantizar un análisis más robusto.

A pesar de su efectividad en la detección de violencia verbal, el modelo podría beneficiarse de futuras optimizaciones, como la integración de técnicas avanzadas de análisis de contexto, la extensión a otras redes sociales (como Twitter o Instagram) y la incorporación de algoritmos que permitan detectar la intención y tono del mensaje. Estas mejoras fortalecerían la capacidad de la

herramienta para identificar discursos violentos con mayor precisión, contribuyendo así a la moderación automatizada del contenido en plataformas digitales.

Aguirre, Mena, Alves y Chávez (2023) desarrollaron una herramienta basada en inteligencia artificial con el propósito de detectar discursos de odio en línea dirigidos a comunicadores. Al comparar la herramienta digital creada en el presente estudio con la plataforma *Attack Detector*, implementada por la Asociación Brasileña de Periodismo de Investigación (Abraji) y la organización mexicana Data Crítica, se pueden identificar tanto similitudes como diferencias significativas en cuanto al diseño y la aplicación de ambas soluciones tecnológicas para la detección de contenido violento en entornos digitales.

En términos de enfoque, ambas herramientas emplean técnicas avanzadas de PLN y aprendizaje automático para clasificar automáticamente los comentarios en categorías como violentos o no violentos. No obstante, mientras que "Attack Detector" se centra en el monitoreo del discurso de odio dirigido a periodistas y activistas en redes sociales como Twitter, la presente investigación orienta su análisis hacia la detección de patrones de violencia verbal en los comentarios de YouTube. Esta diferencia en las plataformas de análisis es relevante, ya que el tipo de interacción, la dinámica del contenido y la estructura de los comentarios varían significativamente entre Twitter y YouTube. En YouTube, los comentarios suelen estar vinculados a videos específicos, mientras que en Twitter predominan conversaciones en hilos y respuestas directas, lo que influye en la forma en que se manifiestan los discursos agresivos.

Otra diferencia significativa radica en la visualización y generación de informes. Mientras que "Attack Detector" se enfoca en la detección automatizada y análisis a gran escala de ataques dirigidos a periodistas, la herramienta del presente estudio incorpora gráficos interactivos y reportes detallados para proporcionar a los usuarios una representación visual clara de la distribución de los comentarios analizados. La implementación de gráficos de barras dinámicos, tablas con información estructurada y una clasificación por palabras clave permite no solo identificar la presencia de violencia verbal, sino también comprender su frecuencia y los términos más utilizados en contextos agresivos.

En cuanto a la metodología de extracción de datos, ambas plataformas emplean técnicas de Web Scraping para la recopilación automatizada de comentarios en redes sociales. Sin embargo, en el presente estudio, se aplicó un proceso de limpieza y tokenización basado en DistilBERT, lo que permitió una clasificación más precisa de los comentarios, con una precisión del 94.12% y un F1-Score de 0.928. En contraste, aunque "Attack Detector" también utiliza modelos de NLP, los detalles específicos sobre sus métricas de desempeño no han sido ampliamente documentados, lo que dificulta una comparación directa en términos de exactitud y eficiencia.

Finalmente, una de las principales limitaciones compartidas por ambas herramientas es la dificultad en la interpretación del contexto semántico. En este estudio, se identificó que comentarios irónicos, sarcásticos o con significados implícitos pueden ser clasificados erróneamente. De manera similar, "Attack Detector" enfrenta desafíos en la detección de ataques que dependen del contexto lingüístico y cultural. Ambas investigaciones sugieren la necesidad de mejorar los modelos de análisis con conjuntos de datos más amplios y estrategias de refinamiento semántico, así como la posibilidad de expandir su aplicabilidad a otras plataformas digitales.

En líneas generales, aunque ambos estudios presentan un alto grado de convergencia en el uso de PLN y aprendizaje automático para la detección de discurso violento en redes sociales, la herramienta desarrollada en esta investigación ofrece un enfoque más orientado a la visualización interactiva y análisis detallado de comentarios en YouTube, mientras que "Attack Detector" prioriza el análisis masivo y la detección de ataques dirigidos a comunicadores en Twitter. Estas diferencias destacan la importancia de diseñar herramientas especializadas en función del tipo de plataforma y del objetivo específico de monitoreo del discurso en línea.

5.3 Conclusión General

El presente estudio ha demostrado la efectividad del análisis automatizado de comentarios en redes sociales para la detección de violencia verbal, utilizando técnicas avanzadas de PLN y aprendizaje profundo. A través de la implementación

del modelo DistilBERT, se logró un sistema capaz de identificar patrones de agresión en textos provenientes de plataformas digitales como Facebook y YouTube, proporcionando un enfoque riguroso y estructurado para la clasificación de comentarios violentos y no violentos. Los resultados obtenidos refuerzan el papel de los modelos de NLP en la identificación y mitigación de discursos agresivos en entornos digitales, evidenciando que el uso de modelos preentrenados optimizados permite alcanzar un alto nivel de precisión (94.12%) y exactitud (94.05%), asegurando la confiabilidad de las predicciones.

No obstante, el estudio presenta ciertas limitaciones que deben ser consideradas para su aplicación en escenarios más amplios. En primer lugar, la validación del modelo se realizó sobre un conjunto de datos específico, lo que podría restringir su aplicabilidad a otras plataformas digitales con diferentes dinámicas comunicativas, como Twitter o TikTok. Asimismo, el modelo se entrenó exclusivamente en español, sin considerar la variabilidad dialectal o el impacto de la violencia verbal en distintos contextos lingüísticos y culturales. Además, el enfoque utilizado se basó en una clasificación binaria (violento/no violento), sin evaluar posibles gradaciones de agresividad en los comentarios. Estas limitaciones subrayan la necesidad de continuar explorando estrategias que permitan mejorar la adaptabilidad y precisión del modelo en escenarios más diversos.

A partir de estos hallazgos, se sugieren líneas de investigación futuras que contribuyan a la optimización y generalización del modelo. La implementación de enfoques de aprendizaje no supervisado o modelos de PLN de última generación, como BERT mejorado o GPT, podría mejorar la capacidad del sistema para comprender matices en la violencia verbal. Asimismo, la integración de modelos multimodales que combinen análisis de texto, imágenes y videos permitiría una detección más integral del fenómeno. También sería relevante evaluar el desempeño del modelo en distintas plataformas sociales y explorar la posibilidad de desarrollar herramientas de moderación automatizada que puedan ser implementadas por empresas tecnológicas o entidades reguladoras para la identificación temprana de contenido violento.

En líneas generales, este estudio ha demostrado que el uso de modelos de PLN y técnicas de ciencia de datos es una estrategia viable y efectiva para la detección de violencia verbal en redes sociales. A pesar de las limitaciones identificadas, los resultados obtenidos ofrecen una base sólida para el desarrollo de sistemas automatizados de monitoreo y regulación de contenido digital, contribuyendo a la reducción de la propagación de discursos agresivos en entornos virtuales. Se espera que futuras investigaciones continúen en esta línea, ampliando el alcance del modelo y explorando nuevas metodologías que fortalezcan la identificación y prevención de la violencia verbal en el ecosistema digital.

Conclusiones

Conclusión 1: Los resultados obtenidos en este estudio confirman que el uso de herramientas de ciencia de datos permite detectar con precisión y eficiencia señales de violencia verbal en redes sociales. El modelo desarrollado, basado en DistilBERT y técnicas de Procesamiento de Lenguaje Natural, mostró un desempeño destacado al alcanzar una precisión del 94.12%, una exhaustividad del 91.57%, un F1-Score de 0.928 y una exactitud global del 94.05%. Estas métricas evidencian un equilibrio sólido entre la capacidad del sistema para identificar comentarios violentos y su efectividad en la reducción de errores de clasificación. Asimismo, el índice de Kappa de Cohen (0.88) indica un alto grado de concordancia entre las predicciones del modelo y las etiquetas reales, superando significativamente el azar.

Conclusión 2: El estudio permitió identificar patrones lingüísticos característicos de comentarios con violencia verbal en redes sociales. A través de un proceso sistemático de recolección, etiquetado y preprocesamiento, se conformó una base de datos estructurada con 2343 comentarios extraídos de la plataforma YouTube, clasificados como violentos o no violentos. La aplicación de técnicas de *Web Scraping* facilitó la recopilación de información relacionada con diversas temáticas sociales, asegurando una muestra representativa para el análisis. Los resultados muestran que la violencia verbal en línea suele manifestarse en mensajes breves, con una media de 20.79 palabras, y una alta concentración de términos agresivos. El análisis de frecuencia identificó palabras recurrentes como “corrupto”, “delincuente”, “rata” y “asesino”, lo que refleja una tendencia hacia la descalificación personal y el discurso de odio. Además, se evidenció que el 49.13% de los comentarios analizados contenían violencia verbal, frente a un 50.87% que no presentaban este tipo de contenido, lo que resalta la magnitud del problema en el ecosistema digital actual.

Conclusión 3: El presente estudio ha evidenciado la viabilidad de desarrollar un modelo de análisis de sentimientos basado en técnicas de Procesamiento de Lenguaje Natural (PLN) para la detección de violencia verbal en redes sociales. La implementación de DistilBERT permitió clasificar comentarios

con alta precisión, alcanzando un índice de Kappa de Cohen de 0.88 y una exactitud (Accuracy) del 94.05%. La identificación de patrones lingüísticos agresivos, mediante técnicas de tokenización y codificación, posibilitó una diferenciación efectiva entre expresiones violentas y no violentas. Asimismo, se comprobó que los comentarios violentos contienen términos con carga semántica negativa que se repiten con frecuencia, lo que facilitó el entrenamiento supervisado del modelo. La depuración cuidadosa de los datos y la combinación de métodos de vectorización contribuyeron a mejorar tanto la fiabilidad como la comprensión contextual del sistema.

Conclusión 4: El modelo DistilBERT en la clasificación de comentarios violentos y no violentos ha demostrado un desempeño altamente eficiente, alcanzando valores de precisión del 94.12%, sensibilidad del 91.57% y un F1-Score de 0.928. Estos resultados reflejan un equilibrio sólido entre la correcta identificación de comentarios violentos y la minimización de falsos positivos, consolidando la fiabilidad del modelo para la detección automatizada de violencia verbal en redes sociales. La evaluación a través de la Curva ROC y el cálculo del Área Bajo la Curva (AUC = 0.98) confirman que el modelo posee una alta capacidad de discriminación entre comentarios violentos y no violentos, superando ampliamente el umbral de aleatoriedad. Sin embargo, los errores detectados sugieren la necesidad de continuar optimizando el modelo mediante ajustes en los hiperparámetros y la posible incorporación de estrategias adicionales de preprocesamiento de datos, con el fin de mejorar aún más su sensibilidad y especificidad.

Conclusión 5: El diseño e implementación de una herramienta digital para la detección y análisis de patrones de violencia verbal en redes sociales permitió identificar tendencias lingüísticas agresivas y generar visualizaciones interactivas que facilitan su interpretación. A través de la extracción automatizada de comentarios en YouTube y su clasificación mediante técnicas de Procesamiento de Lenguaje Natural (PLN) y aprendizaje automático, la plataforma desarrollada logró presentar informes detallados y gráficos interactivos que incluyen la categorización del total de comentarios en niveles de violencia baja (menos del 33%), media (menos del 66%) y alta (mayor o igual al 66%), así como gráficos de columnas y

nubes de palabras que resaltan los términos violentos más frecuentes. Estas representaciones contribuyen a una mejor comprensión del fenómeno de la violencia verbal en línea. El sistema también integra reportes visuales en forma de gráficos de barras y tablas dinámicas que reflejan la proporción de comentarios violentos y no violentos dentro del conjunto de datos. Además, incorpora un mecanismo de selección aleatoria de muestras representativas cuando el volumen de datos es elevado, permitiendo un análisis eficiente sin perder información relevante. Estas funcionalidades brindan una representación clara y accesible de la distribución del lenguaje agresivo en los comentarios analizados de la plataforma YouTube, lo que fortalece la utilidad de la herramienta como apoyo para el monitoreo y la toma de decisiones en entornos virtuales. Aunque la categorización de comentarios como violentos o no violentos se realizó únicamente en esta plataforma, los resultados obtenidos sientan las bases para su futura aplicación en otros contextos digitales.

Recomendaciones

Recomendación 1: Desde el punto de vista metodológico, se sugiere ampliar la base de datos en futuras investigaciones, incorporando una mayor diversidad de comentarios mediante la inclusión de múltiples plataformas digitales y variaciones dialectales del español. Esta expansión del corpus podría contribuir a mejorar la capacidad de generalización del modelo y su adaptabilidad a distintos contextos socioculturales. Asimismo, se recomienda validar externamente el modelo con conjuntos de datos independientes para evaluar su desempeño en escenarios distintos a los del entrenamiento. Este enfoque permitiría garantizar una mayor robustez del sistema y reducir posibles sesgos derivados de una muestra limitada.

Recomendación 2: Para garantizar una mayor eficacia del modelo, se recomienda que su entrenamiento se realice utilizando un conjunto de datos amplio, representativo y lingüísticamente diverso, que contemple variaciones del español propias de los entornos digitales. Asimismo, se sugiere incorporar enfoques híbridos que integren métodos estadísticos tradicionales, como TF-IDF y Word2Vec, con modelos de deep learning, lo cual permitiría mejorar la capacidad del sistema para reconocer patrones contextuales en comentarios ambiguos o con significados variables, fortaleciendo así su precisión y adaptabilidad.

Recomendación 3: Para optimizar el rendimiento del modelo DistilBERT en la detección de violencia verbal en redes sociales, se recomienda ajustar los hiperparámetros y optimizar el umbral de clasificación, con el fin de equilibrar la minimización de falsos positivos y falsos negativos. Asimismo, sería conveniente ampliar y diversificar el conjunto de datos incorporando comentarios provenientes de diversas plataformas y contextos lingüísticos, lo cual podría mejorar la capacidad de generalización del modelo. También se sugiere aplicar técnicas avanzadas de preprocesamiento, como la normalización semántica y la expansión de abreviaturas, a fin de mejorar la representación del lenguaje y reducir errores de clasificación. Además, se recomienda comparar el desempeño de DistilBERT con otros modelos de aprendizaje profundo, como RoBERTa o ALBERT, con el propósito de identificar la opción más eficiente. Finalmente, la implementación del

modelo en un entorno de prueba permitiría evaluar su efectividad en situaciones reales y realizar los ajustes necesarios conforme a los requerimientos de moderación automatizada. En conjunto, estas estrategias podrían fortalecer la precisión, sensibilidad y aplicabilidad del sistema en la identificación de patrones de violencia verbal en entornos digitales.

Recomendación 4: Se recomienda ampliar la funcionalidad de la herramienta digital mediante la integración de modelos avanzados de análisis semántico y detección de ironía, lo cual podría mejorar significativamente la precisión en la clasificación de comentarios agresivos. Asimismo, sería conveniente extender su aplicación a múltiples plataformas de redes sociales y desarrollar un sistema de alertas automáticas que permita la monitorización en tiempo real de patrones de violencia verbal. Estas mejoras contribuirían a una identificación más efectiva del discurso violento en entornos digitales y fortalecerían el diseño de estrategias de moderación y prevención en contextos virtuales.

Referencias

Atapattu, T., Rajapakse, R. & Falkner, K. (2020). Detection of cyberbullying and cyber aggression in social media: A survey of techniques and challenges. *Natural Language Processing Journal,* 11(4), 120-140. https://doi.org/10.1016/j.nlpj.2020.03.005

Babu, N. & Kanaga, E. (2022). Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. *SN Computer Science* (3) 74. https://link.springer.com/article/10.1007/s42979-021-00958-1

Badjatiya, P., Gupta, S., Gupta, M. & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. Proceedings of the 26th International Conference on World Wide. https://dl.acm.org/doi/10.1145/3038912.3052591, 1391 - 1399.

Bandura, A. (1977). *Social Learning Theory*. Prentice Hall.

Beck, K. (2001). *Extreme Programming Explained: Embrace Change.* (2DA Edición). Boston, MA: Addison-Wesley. https://ptgmedia.pearsoncmg.com/images/9780321278654/samplepages/9780321278654.pdf

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Cabrera, P. (2023). Nueva organización de los diseños de investigación. *South American Research Journal*, 3(1), 37–51. https://www.sa-rj.net/index.php/sarj/article/view/37

Campillo, S. (2019). Propuesta de clasificación de actos verbales violentos en las redes sociales. Revista E-Aesla, 5, 199-207. https://cvc.cervantes.es/lengua/eaesla/pdf/05/20.pdf,

Campillo, S. (2022). Lingüística de corpus y análisis de la violencia verbal en las redes sociales. *Recerca en humanitats 2020,* 27-36. Universitat Rovira i Virgili.

https://www.researchgate.net/publication/357869305_Linguistica_de_corpus_y_analisis_de_la_violencia_verbal_en_las_redes_sociales

Carvajal, D. (2023). *Modelos de Deep learning para la clasificación de textos en los objetivos de desarrollo sostenible*. Universidad de los Andes, Departamento de Ingeniería de Sistemas y Computación. Bogotá D.C., Colombia. https://repositorio.uniandes.edu.co/server/api/core/bitstreams/4a490d70-0f30-4679-9dc8-3694b96933e8/content

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217-1230. https://dl.acm.org/doi/pdf/10.1145/2998181.2998213

Citron, D. (2014). Hate crimes in cyberspace. *Harvard University Press.* https://scholarlycommons.law.case.edu/jolti/vol6/iss1/3/

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37-46.

Cortez, A., Vega, H. & Pariona, J. (2013). Procesamiento de lenguaje natural. Primer encuentro de Grupos de investigación sobre Procesamiento del lenguaje, *Revista PLN*, 51. file:///C:/Users/PILAR/Downloads/PLN_51.pdf

Crane, D. (2019). *Text Mining and Analysis in Social Media: A Practical Guide*. Nueva York, EE. UU.: Oxford University Press, https://www.amazon.com/Text-Mining-Analysis-Practical-Examples/dp/161290551X.

Cueva, T., Jara, O., Arias, J., Flores, F. & Balmaceda, C. (2023). *Métodos mixtos de investigación para principiantes.* INUDI Perú. https://editorial.inudi.edu.pe/index.php/editorialinudi/catalog/download/119/161/190?inline=1

Cunningham, W. (2015). *The New Digital Economy: The Future of Work*. Nueva York, NY: N.H. International Publishers, https://www.amazon.com/New-Digital-Economy-Future-Work/dp/1138461376.

Davenport, T. y Kirby, J. (2020). *The AI Advantage: How Robotic Process Automation is Transforming* Business. Boston, EE. UU.: Harvard Business Review Press.

Deng, L. & Liu, Y. (2018). *Deep learning in natural language processing.* Springer.

Fairclough, N. (1992). Discourse and Social Change. Polity Press.

Galván, J. (2024, junio). Linkedln. La violencia verbal en redes sociales: Un Golpe Directo a la Salud Mental. *Lindkelin* https://www.linkedin.com/pulse/la-violencia-verbal-en-redes-sociales-un-golpe-directo-josefa-galv%C3%A1n-6joxc/

Hatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G. & Vakali, A. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)* 13(3), 1-51. https://dl.acm.org/doi/10.1145/3343484

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la Investigación.* (Sexta Edición). McGraw-Hill / Interamericana Editores, S.S. de C.V.

IA University. (s.f.). La inteligencia artificial y el análisis de sentimiento en las redes sociales para el marketing digital. https://ia.university/actualizar-archivos-dll/la-inteligencia-artificial-y-el-analisis-de-sentimiento-en-las-redes-sociales-para-el-marketing-digital/

International Republican Institute. (2016, diciembre.). *Violencia en redes sociales: Un fenómeno que debemos prevenir.* https://www.iri.org/news/violencia-en-redes-sociales-un-fenomeno-que-debemos-prevenir/

Khan, M., Malik, M. & Nadeem, A. (2024). Detection of violence incitation expressions in Urdu tweets using convolutional neural network. *Expert Systems with Applications* Volume 245. https://www.sciencedirect.com/science/article/pii/S0957417424000393

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Londres. Reino Unido: *SAGE Publications,* https://methods.sagepub.com/book/the-data-revolution.

Kont, J. (2016, diciembre). Violencia en redes sociales: Un fenómeno que debemos prevenir. *International Republican Institute*. https://www.iri.org/news/violencia-en-redes-sociales-un-fenomeno-que-debemos-prevenir/

Kont, José. (2016, diciembre). Violencia en Redes Sociales: Un fenómeno que debemos prevenir. *International Republican Institute.* https://www.iri.org/news/violencia-en-redes-sociales-un-fenomeno-que-debemos-prevenir/

Kuz A., Falco M. & Giandini R. (2016). Análisis de redes sociales: un caso práctico. *Revista Computación y Sistemas* 20(1),89-106. https://gc.scalahed.com/recursos/files/r161r/w25553w/Analisis_redes_sociales.pdf,

Lacunza, A., Contini, E., Caballero, S. & Mejail, S. (2019). Agresión en las redes y adolescencia: Estado actual en América Latina desde una perspectiva bibliométrica. *Investigación y Desarrollo,* 27(2), 6-32. Fundación Universidad del Norte. https://www.redalyc.org/journal/268/26864302001/html/

LISA. (2022, enero). ¿Qué hay detrás del origen del PLN? *LISA Insurtech.* https://lisainsurtech.com/que-hay-detras-del-origen-del-pln/

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologi, California: Morgan & Claypool Publishers, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Loshchilov, I. & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.* https://arxiv.org/abs/1711.05101

Magoga, A. (2018). *Prevención de la violencia juvenil a través de las TIC en El Salvador, Honduras y Guatemala.* Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000368756.locale=en

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA, Estados Unidos: MIT Press, https://nlp.stanford.edu/fsnlp/.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv:1301.3781. https://arxiv.org/pdf/1301.3781

Moya, D. y Majó, J. (2017). ANÁLISIS DE COMENTARIOS EN REDES SOCIALES PARA MEJORAR LA REPUTACIÓN ONLINE HOTELERA. Turismo y Sociedad, 169-190.

Ñaupas et al. (2018). Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis. (5° ed., pp 140-141). Bogotá: Ediciones de la U. http://www.biblioteca.cij.gob.mx/archivos/materiales_de_consulta/drogas_de_abuso/articulos/metodologiainvestigacionnaupas.pdf

Oficina para la Salud de la Mujer [OASH (2023). Abuso emocional y verbal. Departamento de Salud y Servicios Humanos de los Estados Unidos. https://espanol.womenshealth.gov/relationships-and-safety/other-types/emotional-and-verbal-buse

Organización de las Naciones Unidas [ONU] (2023, diciembre). Más del 75% de la población mundial tiene un teléfono celular y más del 65% usa el internet. Noticias ONU. https://news.un.org/es/story/2023/12/1526712

Organización de las Naciones Unidas para la Educación [UNESCO]. (2022). *Countering online hate speech.* https://unesdoc.unesco.org/ark:/48223/pf0000233231

Organización Mundial de la Salud [OMS] (2022). Informe mundial sobre salud mental: Transformar la salud mental para todos. Organización Mundial de la Salud. https://iris.who.int/bitstream/handle/10665/356118/9789240051966-spa.pdf?sequence=1

Papacharissi, Z. (2004). "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." New Media & Society, 6(2), 259-283. https://journals.sagepub.com/doi/abs/10.1177/1461444804041444

Pardo, A., Ruiz, M. & San Martín, R. (2012). Análisis de datos en ciencias sociales de la salud. (1° ed., pp 35-36). Síntesis. España.

Patchin, J. & Hinduja, S. (2020). Cyberbullying: Identification, prevention, and response. *Cyberbullying Research Center.* https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2020.pdf?utm_source

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Peña, P. (2021). Entre analogías y metáforas: El debate sobre la moderación de contenidos en las redes sociales. *Revista de las Cortes Generales,* (111), 265-311. https://doi.org/10.33426/rcg/2021/111/1614

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Jo*urnal of Machine Learning Technologies,* 2(1), 37-63. https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation

Quiroz, Y. (2014). *Las redes sociales como herramientas del periodismo digital.* Facultad de Ciencias de la Comunicación, Turismo y Psicología, 28, 279-303.

Riofrío, D. & Cumba, P. (2018). Predicción de ataques de cyber bullying mediante técnicas de aprendizaje profundo apoyándose en un corpus de entrenamiento para la clasificación de texto en español. [Tesis de Maestría, Universidad Internacional SEK]. Repositorio Institucional UISEK. https://repositorio.uisek.edu.ec/handle/123456789/3224

Robert, M. (2011). *Clean Code: A Handbook of Agile Software Craftsmanship*. Upper Saddle River, NJ: Prentice Hall. https://www.amazon.com/-/es/Robert-C-Martin/dp/8441532109.

Sánchez, O. (2023, diciembre). *Análisis de sentimientos: Un viaje a través de las emociones (Parte I).* LinkedIn. https://es.linkedin.com/pulse/an%C3%A1lisis-de-sentimientos-un-viaje-trav%C3%A9s-las-parte-i-octavio-s%C3%A1nchez-zoqwc#:~:text=En%20la%20d%C3%A9cada%20de%201980,en%20el%20an%C3%A1lisis%20de%20sentimientos

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.* https://arxiv.org/pdf/1910.01108

Schmidt, A. & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.* 1–10). Association for Computational Linguistics. https://aclanthology.org/W17-1101/?utm_source

Silva, D., Sousa, P. & Pereira, J. (jan-feb., 2019). A aplicação da técnica de análise de sentimento em mídias sociais como instrumento para as práticas da gestão social em nível governamental. *Revista de administración pública* 53(1). https://www.scielo.br/j/rap/a/GD3F8HdkQKGSHy8zzV8w9Ys/#

Simplified. (s.f.). Facebook Comment. Simplified. Recuperado el 4 de febrero de 2025. https://simplified.com/es/social-media-glossary/facebook-comment

Smith, J., Petrovic, P., Rose, M., De Souz, C., Muller, L., Nowak, B. & Martinez, J. (2023). Placeholder Text: A Study. *The Journal of Citation Styles,* 3. https://rediech.org/ojs/2017/index.php/recie/article/view/2012/2094

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior,* 7(3), https://www.researchgate.net/publication/8451443_The_Online_Disinhibition_Effect

Sussman, G. & Steele, G. (1975). The Lambda Calculus: Its Syntax and Semantics. *Cambridge, MA: MIT Press*, https://www.amazon.com/Calculus-Semantics-Studies-Foundations-Mathematics/dp/0444875085.

Van Hoecke, H. y Pauwels, J. (2018). Web scraping for social sciences: A comprehensive guide. Londres, Reino Unido: SAGE Publications. https://link.springer.com/book/10.1007/978-3-030-36409-7.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. & Polosukhin, I. (2017). *Attention is all you need. Advances in Neural Information Processing Systems, 30*. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Waseem, Z., Davidson, T., Warmsley, D. & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *En Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). Association for Computational Linguistics. https://aclanthology.org/W17-3012/?utm_source

Willcocks, L., Lacity, M. & Craig, A. (2015). Robotic Process Automation: The Next Transformation Lever for Shared Services. *London School of Economics.* https://www.researchgate.net/publication/273871891_Intervalnye_i_predelnye_r_ph-harakteristiki_funkcii_Vejerstrassa.

Wirth, R. & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.* Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39. https://link.springer.com/chapter/10.1007/978-3-540-39804-2_9

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A. & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language *Processing: System Demonstrations,* 38-45. https://www.aclweb.org/anthology/2020.emnlp-demos.6/

Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine,* 13(3), 55-75. https://arxiv.org/pdf/1708.02709

Anexos

Anexo 1. Matriz de consistencia

Título: “Detección de señales de violencia en redes sociales: Un enfoque desde la ciencia de datos”

Pregunta general	Preguntas específicas	Objetivo general	Objetivos específicos	Variables / Categorías	Dimensiones / Subcategorías	Enfoque, tipo y diseño	Población y muestra	Técnicas e instrumentos
¿Cómo se pueden detectar señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos?	1. ¿Qué características y patrones de lenguaje están presentes en comentarios y publicaciones que contienen violencia verbal en redes sociales?	Identificar señales de violencia verbal en comentarios y publicaciones de redes sociales mediante herramientas de ciencia de datos.	1. Identificar y analizar las características y patrones de lenguaje en comentarios de redes sociales que contienen violencia verbal, mediante la extracción, etiquetado y preprocesamiento de un conjunto de datos.	**Variable 1**: Comentarios y publicaciones de redes sociales.	* Palabras clave relacionadas con violencia verbal. * Frecuencia de comentarios violentos.	**Enfoque**: Mixto: Cuantitativo y cualitativo. **Tipo y alcance:** Aplicado y Descriptivo **Diseño:** Secuencial exploratorio, no experimental y observacional.	**Población:** Comentarios en la red social YouTube y Facebook Perú 2024. **Muestra:** 9556 comentarios extraídos de la plataforma YouTube que contengan comentarios violentos y no violentos	- Técnicas de automatización de procesos robóticos (RPA) y de preprocesamiento de datos. - Análisis de sentimientos mediante *machine learning* o *deep learning* - Evaluación de métricas como precisión, sensibilidad y especificidad - Visualización de informes y ventanas emergentes utilizando herramientas digitales.
	2. ¿Cómo se puede desarrollar un modelo de análisis de sentimientos basado en técnicas de procesamiento de lenguaje natural (PLN) para detectar señales de violencia verbal en redes sociales?		2. Desarrollar un modelo de análisis de sentimientos para detectar violencia verbal con técnicas de NLP.					
	3. ¿Qué nivel de precisión, sensibilidad y especificidad puede alcanzar el modelo de análisis de sentimientos para detectar violencia verbal?		3. Evaluar y optimizar el rendimiento del modelo con métricas de clasificación.	**Variable 2**: Categoría del sentimiento de violencia verbal.	* Nivel de violencia en el texto. * Precisión y rendimiento del modelo.			
	4. ¿Qué tipo de informes y visualizaciones podrían ser útiles para identificar tendencias y patrones de violencia verbal en redes sociales, y cómo pueden implementarse alertas automáticas?		4. Diseñar una herramienta digital que genere informes y ventanas emergentes que muestren patrones de violencia verbal.					

Anexo 2. Cripts Python para la descripción de los comentarios.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from scipy.stats import  kurtosis , skew
import re
```

```
def buscar_palabras_en_oraciones(palabras, oraciones):
    contiene_palabra = []
    ocurrencias = []
    # Diccionario para contar frecuencia de cada palabra
    frecuencia_palabras  =  {str(palabra).lower():  0  for  palabra  in
palabras}

    for oracion in oraciones:
        # Inicializar el contador de ocurrencias
        contador_ocurrencias = 0
        oracion = str(oracion).lower()

        # Recorrer cada palabra de la lista
        for palabra in palabras:
            palabra = str(palabra).lower()
            # Contar ocurrencias de esta palabra específica
            ocurrencias_palabra = oracion.count(palabra)
            # Actualizar el contador total
            contador_ocurrencias += ocurrencias_palabra
            # Actualizar el contador específico de la palabra
            frecuencia_palabras[palabra] += ocurrencias_palabra

        # Si se encontraron palabras, agrega 1, si no, 0
        contiene_palabra.append(1 if contador_ocurrencias > 0 else 0)
        ocurrencias.append(contador_ocurrencias)

    # Ordenar el diccionario por frecuencia de mayor a menor
    palabras_ordenadas = dict(sorted(frecuencia_palabras.items(),
                                     key=lambda x: x[1],
                                     reverse=True))

    return contiene_palabra, ocurrencias, palabras_ordenadas
```

```python
def etiquetas_L(texto):
    if not isinstance(texto, str):  # Verificar si el valor es una cadena
        return ''   # Si no es un texto, devolver un string vacío o un
valor por defecto
    texto_limpio = re.sub(r'[^a-zA-Z0-9áéíóúÁÉÍÓÚüÜñÑ\s]', '', texto)
    # Reemplazar múltiples espacios consecutivos por un solo espacio
    texto_limpio = re.sub(r'\s+', ' ', texto_limpio).strip()

    # Eliminar todo lo que está después de un punto (incluyendo el punto)
    texto_limpio = re.sub(r'\..*', '', texto_limpio).strip()
    texto_limpio = texto_limpio.replace(" ", "")
    return texto_limpio
```

```python
def limpiar_comentarios(texto):
    if not isinstance(texto, str):  # Verificar si el valor es una cadena
        return ''   # Si no es un texto, devolver un string vacío o un
valor por defecto
    texto_limpio = re.sub(r'[^a-zA-Z0-9áéíóúÁÉÍÓÚüÜñÑ\s]', '', texto)
    # Reemplazar múltiples espacios consecutivos por un solo espacio
    texto_limpio = re.sub(r'\s+', ' ', texto_limpio).strip()

    # Eliminar todo lo que está después de un punto (incluyendo el punto)
    texto_limpio = re.sub(r'\..*', '', texto_limpio).strip()
    return texto_limpio


def contar_palabras(Texto):
    numero_text = str(Texto).split()
    return len(numero_text)
```

```
# Ejemplo de uso:
comments_Youtube = pd.read_excel("datos_resampled3a.xlsx")
etiquetas = pd.read_excel("etiquetas (2) (1) (1).xlsx")

etiquetas['Palabras'] = etiquetas['Palabras'].apply(etiquetas_L)
etiquetas.drop_duplicates(subset=['Palabras'], inplace=True)


comments_Youtube['Comentario Limpio'] =
comments_Youtube['Comentario'].apply(limpiar_comentarios)
comments_Youtube['Numero palabras'] = comments_Youtube['Comentario
Limpio'].apply(contar_palabras)

comments_Youtube['Is_Ira'] =
buscar_palabras_en_oraciones(etiquetas['Palabras'].values ,
comments_Youtube['Comentario'].values)[0]
comments_Youtube['Ocurrencia'] =
buscar_palabras_en_oraciones(etiquetas['Palabras'].values ,
comments_Youtube['Comentario'].values)[1]
palabras_frecuentes =
buscar_palabras_en_oraciones(etiquetas['Palabras'].values ,
```

```
from transformers import DistilBertTokenizer, DistilBertForSequenceClassification
from torch.utils.data import DataLoader
from sklearn.model_selection import train_test_split
import torch
import pandas as pd
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
```

```
# Cargar el tokenizer
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')

# Tokenización de los comentarios
def tokenizar_texto(texto):
    return tokenizer(texto,
```

```
df_palabras_frecuentes =
pd.DataFrame(list(palabras_frecuentes.items()), columns=['Palabra',
'Frecuencia'])
df_palabras_frecuentes.drop_duplicates(subset=['Palabra'] , inplace=
True)
```

```
palabras_frecuentes = {k.replace('\n', ' ') : v for k, v in
palabras_frecuentes.items()}


# Crear la nube de palabras
nube = WordCloud(width=800, height=400,
background_color='white').generate_from_frequencies(palabras_frecuente
s)

# Mostrar la nube de palabras
plt.figure(figsize=(15, 5))
plt.imshow(nube, interpolation='bilinear')
plt.axis('off')  # No mostrar los ejes
```

```
n_clases = 12

# Calculamos la amplitud (redondeada)
amplitud = round((max_valor - min_valor) / n_clases)

# Creamos los límites de clase (asegurando que sean enteros)
limites = np.arange(min_valor, max_valor + amplitud + 1, amplitud)

# Creamos la tabla de frecuencias
frecuencias    =    pd.cut(clean['Numero    palabras'],    bins=limites,
right=False).value_counts().sort_index()

# Creamos el DataFrame con la tabla de frecuencias
tabla_frecuencia = pd.DataFrame({
    'Límite Inferior': limites[:-1],
    'Límite Superior': limites[1:],
    'Frecuencia': frecuencias,
    'Frecuencia Relativa': frecuencias / len(clean),
})

# Añadimos las frecuencias acumuladas
tabla_frecuencia['Frecuencia                Acumulada']                =
tabla_frecuencia['Frecuencia'].cumsum()
tabla_frecuencia['Frecuencia        Relativa        Acumulada']        =
tabla_frecuencia['Frecuencia Relativa'].cumsum()

# Mostrar la nube de palabras
plt.figure(figsize=(15, 5))
plt.imshow(nube, interpolation='bilinear')
plt.axis('off')  # No mostrar los ejes
plt.show()
```

```
plt.figure(figsize=(6,4))
sns.histplot(clean['Ocurrencia'], bins=21 , kde=True)
plt.ylabel("Frecuencia" , fontsize=9, fontweight='bold')
plt.xlabel("Numero de palabras con \níra por comentario" , fontsize=9,
fontweight='bold')
plt.grid(axis='y', linestyle='--', linewidth=0.5)  # Línea discontinua e
intensidad reducida
plt.tick_params(axis='y', which='both', left=False)  # Quita las marcas
en el eje Y si es necesario

# Mostrar el gráfico
plt.show()
```

```
# Crear el gráfico circular (pie chart)
plt.figure(figsize=(6, 6))

# Graficar el pie chart con borde, sombra y colores personalizados
Is_ira['Frecuencia'].plot.pie(autopct='%1.1f%%', labels=['No violentas'
, 'Violentas'], startangle=90,
                              colors=['#2e86c1',          '#d6eaf8'],
fontsize=14,
                              wedgeprops={'edgecolor':       'black',
'linewidth': 1, 'linestyle': 'solid'})

# Ajustar la visibilidad y el estilo de los textos
plt.axis('off')  # Desactivar los ejes
```

Anexo 3. Scripts Python para el entrenamiento y evaluación del modelo

```python
from transformers import DistilBertTokenizer, DistilBertForSequenceClassification
from torch.utils.data import DataLoader
from sklearn.model_selection import train_test_split
import torch
import pandas as pd
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import torch
from torch.utils.data import DataLoader
from transformers import DistilBertTokenizer, DistilBertForSequenceClassification
from sklearn.model_selection import train_test_split
from torch.optim import AdamW

import torch
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, roc_curve, auc
from sklearn.preprocessing import LabelBinarizer
```

```python
df = pd.read_excel('/content/datos_resampled3a.xlsx')
df.dropna(subset=['Comentario'], inplace=True)
df.reset_index(drop=True, inplace=True)
```

```
# Cargar el tokenizer
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')

# Tokenización de los comentarios
def tokenizar_texto(texto):
    return tokenizer(texto,
                     padding=True,          # Agrega padding a las frases más cortas
                     truncation=True,       # Trunca las frases más largas
                     max_length=512,        # Define la longitud máxima de la secuencia
                     return_tensors="pt",   # Devuelve los resultados como tensores de
PyTorch
                     return_attention_mask=True)  # Asegúrate de retornar el
attention mask también


# Asegúrate de que 'df' esté bien cargado antes de esta línea (df debe ser un DataFrame)
df['tokenized'] = df['Comentario'].apply(tokenizar_texto)

# Etiquetas
y = df['Is_Ira']

# Dividir el conjunto de datos en entrenamiento y prueba
train_texts, test_texts, train_labels, test_labels = train_test_split(df['Comentario'],
y, test_size=0.2)

# Tokenización de los textos
train_encodings = tokenizer(list(train_texts), truncation=True, padding=True,
return_tensors="pt", return_attention_mask=True)
test_encodings = tokenizer(list(test_texts), truncation=True, padding=True,
return_tensors="pt", return_attention_mask=True)

# Convertir etiquetas a tensores
train_labels_tensor = torch.tensor(train_labels.values, dtype=torch.long)
test_labels_tensor = torch.tensor(test_labels.values, dtype=torch.long)

# Crear los datasets de entrenamiento y prueba
train_dataset = TensorDataset(train_encodings['input_ids'],
train_encodings['attention_mask'], train_labels_tensor)
test_dataset = TensorDataset(test_encodings['input_ids'],
test_encodings['attention_mask'], test_labels_tensor)

# Crear los DataLoaders
train_dataloader = DataLoader(train_dataset, batch_size=16, shuffle=True)
test_dataloader = DataLoader(test_dataset, batch_size=16)

# Verificar si CUDA está disponible y usarla
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
print(f"Usando dispositivo: {device}")

# Cargar el modelo DistilBERT
model = DistilBertForSequenceClassification.from_pretrained('distilbert-base-uncased',
num_labels=2)
model.to(device)  # Mover el modelo a la GPU si está disponible

# Configuración del optimizador
optimizer = AdamW(model.parameters(), lr=1e-5)
```

```
# Entrenamiento del modelo
model.train()
for epoch in range(100):  # Ajusta el número de épocas si es necesario
    for batch in train dataloader:
        optimizer.zero grad()

        input ids, attention mask, labels = batch
        input ids = input ids.squeeze(1).to(device)  # Elimina la dimensión extra y
mueve a GPU/CPU
        attention_mask = attention_mask.squeeze(1).to(device)  # Mueve el attention_mask
a la GPU/CPU
        labels = labels.to(device)

        # Realiza el paso hacia adelante
        outputs = model(input_ids, attention_mask=attention_mask, labels=labels)

        loss = outputs.loss
        loss.backward()
        optimizer.step()

    print(f"Epoch {epoch + 1} completado con pérdida: {loss.item()}")

# Evaluación del modelo
model.eval()
correct = 0
total = 0
with torch.no_grad():
    for batch in test dataloader:
        input ids, attention mask, labels = batch
        input_ids = input_ids.squeeze(1).to(device)  # Elimina la dimensión extra y
mueve a GPU/CPU
        attention mask = attention mask.squeeze(1).to(device)  # Mueve el attention mask
a la GPU/CPU
        labels = labels.to(device)

        # Realiza las predicciones
        outputs = model(input_ids, attention_mask=attention_mask)
        logits = outputs.logits
        predictions = torch.argmax(logits, dim=-1)

        correct += (predictions == labels).sum().item()
        total += labels.size(0)

accuracy = correct / total
print(f"Precisión en el conjunto de prueba: {accuracy:.4f}")
```

```
# Evaluación del modelo
model.eval()
correct = 0
total = 0

all_predictions = []
all_labels = []
all_probs = []  # Para almacenar las probabilidades predichas

# Asegúrate de que los tensores estén en el dispositivo correcto
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)  # Mover el modelo al dispositivo

with torch.no_grad():
    for batch in test_dataloader:
        input_ids, attention_mask, labels = batch  # Desempaquetar los 3 elementos

        # Mover los tensores al dispositivo adecuado
        input_ids = input_ids.squeeze(1).to(device)  # Elimina la dimensión extra y mueve a GPU/CPU
        attention_mask = attention_mask.squeeze(1).to(device)  # Mueve el attention_mask a la
GPU/CPU
        labels = labels.to(device)  # Mover las etiquetas a GPU/CPU

        outputs = model(input_ids, attention_mask=attention_mask)
        logits = outputs.logits
        predictions = torch.argmax(logits, dim=-1)
        probs = torch.nn.functional.softmax(logits, dim=-1)  # Probabilidades para cada clase

        # Almacena las predicciones, probabilidades y las etiquetas reales
        all_predictions.extend(predictions.cpu().numpy())
        all_labels.extend(labels.cpu().numpy())
        all_probs.extend(probs.cpu().numpy())

        correct += (predictions == labels).sum().item()
        total += labels.size(0)

# Calcular la precisión
accuracy = correct / total
print(f"Precisión en el conjunto de prueba: {accuracy:.4f}")

# Calcular la matriz de confusión
cm = confusion_matrix(all_labels, all_predictions)

# Mostrar la matriz de confusión
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=["Clase 0", "Clase 1"],
yticklabels=["Clase 0", "Clase 1"])
plt.xlabel('Predicciones')
plt.ylabel('Etiquetas Reales')
plt.title('Matriz de Confusión')
plt.show()

# Curva ROC y AUC
# Binarizar las etiquetas reales (para el caso de clasificación binaria)
lb = LabelBinarizer()
all_labels_bin = lb.fit_transform(all_labels)

# Calcula la curva ROC y AUC
fpr, tpr, thresholds = roc_curve(all_labels_bin, [p[1] for p in all_probs])  # Usamos las
probabilidades de la clase positiva
roc_auc = auc(fpr, tpr)

# Graficar la curva ROC
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='blue', lw=2, label=f'Curva ROC (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='gray', linestyle='--')  # Línea diagonal (aleatoria)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Tasa de Falsos Positivos')
plt.ylabel('Tasa de Verdaderos Positivos')
plt.title('Curva ROC')
plt.legend(loc='lower right')
plt.show()
```

```
# 1. Cargar el archivo Excel con el nuevo conjunto de datos
file_path = '/content/frases.xlsx'  # Reemplaza con la ruta de tu nuevo archivo
Excel
df nuevo = pd.read excel(file path)

# Verifica las primeras filas para asegurarte de que cargaste bien los datos
print(df nuevo.head())

# 2. Cargar el modelo y el tokenizer entrenados
device = torch.device("cuda" if torch.cuda.is available() else "cpu")

# Cargar el tokenizer y el modelo entrenado
model =
DistilBertForSequenceClassification.from_pretrained('/content/path_to_save_model',
num labels=2)  # Reemplaza con la ruta donde guardaste el modelo
tokenizer =
DistilBertTokenizer.from_pretrained('/content/path_to_save_tokenizer')  #
Reemplaza con la ruta donde guardaste el tokenizer

model.to(device)  # Mover el modelo al dispositivo adecuado
model.eval()  # Modo evaluación

# 3. Tokenizar las nuevas oraciones
def tokenizar_texto(texto):
    return tokenizer(texto, padding=True, truncation=True, return tensors="pt",
max_length=512)

# 4. Realizar predicciones con el modelo
def predecir(texto_tokenizado):
    input ids = texto tokenizado['input ids'].to(device)
    attention mask = texto tokenizado['attention mask'].to(device)
    with torch.no_grad():
        outputs = model(input_ids, attention_mask=attention_mask)
        logits = outputs.logits
        predicciones = torch.argmax(logits, dim=-1)  # Predicción final (clase 0 o
1)
    return predicciones.item()

# Tokenizar las oraciones del nuevo conjunto de datos
tokenized data nuevo = df nuevo['Comentario'].apply(tokenizar texto)

# Realizar predicciones para todas las oraciones
df nuevo['Predicciones'] = tokenized data nuevo.apply(predecir)

# 5. Ver los resultados
print(df nuevo[['Comentario', 'Predicciones']])

# 6. Guardar las predicciones en un nuevo archivo Excel (opcional)
df_nuevo.to_excel('predicciones_nuevas_resultado.xlsx', index=False)
```