UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Análisis de Sentimiento y Clasificación de Texto para la Detección Automática de Acosos y Amenazas Mediante Inteligencia Artificial

Sentiment Analysis and Text Classification for Automatic Detection of Harassment and Threats Using Artificial Intelligence

Autores

* ***Kevin Alexander Mendoza Campoverde***
kmendoza7@utmachala.edu.ec

Javier Valentin Hurtado Gonzalez
jhurtado6@utmachala.edu.ec

Rodrigo Fernando Morocho Román
rmorocho@utmachala.edu.ec

Wilmer Braulio Rivas Asanza
wrivas@utmachala.edu.ec

Universidad Técnica de Machala, Facultad de Ingeniería Civil, Machala, El Oro, Ecuador.
*Autor para correspondencia

Comó citar el artículo:
Mendoza Campoverde, K.A., Hurtado Gonzalez, J.V., Morocho Román, R.F. & Rivas Asanza, W.B. (2025). Análisis de Sentimiento y Clasificación de Texto para la Detección Automática de Acosos y Amenazas Mediante Inteligencia Artificial. *Informática y Sistemas, 9*(1), 82–92. https://doi.org/10.33936/isrtic.v9i1.7470

Enviado: 22/04/2025
Aceptado: 20/05/2025
Publicado: 21/05/2025

Resumen

El presente trabajo muestra una comparación entre dos modelos de inteligencia artificial para la detección de lenguaje agresivo en redes sociales entre un modelo tradicional de clasificación de texto y un modelo basado en redes neuronales profundas. Se utilizaron dos enfoques principales: regresión logística utilizando vector TF-IDF y un modelo basado en BERT adaptado para procesamiento de lenguaje natural. En cuanto a la metodología se aplicó CRISP-DM, abordando desde la preparación de los datos hasta la parte final que es la evaluación de los modelos. Se presentaron balances en el conjunto de datos, el cual se corrigió usando la técnica SMOTE. La evaluación de modelos nos demostró que BERT alcanzó mejores métricas de rendimiento con una medida F1 promedio de 0.93 en comparación a la regresión logística que presentó un 0.83. Las métricas junto con la revisión de errores de clasificación ayudaron a observar de forma más clara en qué aspectos cada enfoque presentaba fortalezas o mostraba limitaciones. En síntesis, los resultados obtenidos manifiestan que BERT ofrece ventajas importantes para la tarea de moderación de contenido en redes sociales y además se pudo confirmar que el preprocesamiento adecuado y el balanceo de los datos son factores clave para mejorar el rendimiento en problemas relacionados con la clasificación de texto.

Palabras clave: Ciberacoso; Clasificación de texto; BERT; Regresión logística; Redes sociales

Abstract

This paper shows a comparison between two artificial intelligence models for the detection of aggressive language in social networks between a traditional text classification model and a model based on deep neural networks. Two main approaches were used: logistic regression using TF-IDF vector and a BERT-based model adapted for natural language processing. As for the methodology, CRISP-DM was applied, addressing from data preparation to the final part which is the evaluation of the models. Balances were presented in the data set, which was corrected using the SMOTE technique. The model evaluation showed us that BERT achieved better performance metrics with an average F1 measure of 0.93 compared to logistic regression which presented a 0.83. The metrics together with the review of classification errors helped to observe more clearly in which aspects each approach presented strengths or showed limitations. In summary, the results obtained show that BERT offers significant advantages for the task of content moderation in social networks, and it was also possible to confirm that proper preprocessing and data balancing are key factors to improve performance in problems related to text classification.

Keywords: Cyberbullying; Text classification; BERT; Logistic regression; Social media

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

1. Introducción

¿Cómo ha transformado Internet la forma en que las personas se comunican en las últimas dos décadas? Más allá de ser un ecosistema global donde la información fluye sin fronteras, también ha traído consigo una serie de desafíos en torno a la seguridad y el bienestar de los usuarios en plataformas digitales. Chérrez y Ávila-Pesantez (2021) , señalan que las redes sociales han evolucionado hasta convertirse en espacios donde la vulnerabilidad de los usuarios se multiplica exponencialmente. Dentro de los problemas que existen uno de estos es alarmante el cual es el ciberacoso, el cual hoy en día este problema supone un grave riesgo para la salud emocional y física de un gran número de usuarios de Internet. Gracias al anonimato informático y el fácil acceso a las plataformas digitales el ciberacoso ha aumentado dando cabida a expresiones discriminatorias y violentas. Según el Observatorio Nacional de Tecnología y Sociedad (2022) : El 46% declara haber sufrido acoso en alguna ocasión, lo que ha generado consecuencias psicológicas severas, desde cuadros de ansiedad y depresión hasta ideación suicida. Frente a la magnitud de este problema, los métodos tradicionales de moderación se han quedado cortos. La cantidad de contenido generado en plataformas como Twitter, Facebook e Instagram hace prácticamente imposible la intervención humana exhaustiva. Mientras un moderador humano apenas puede analizar 100 publicaciones por hora, un sistema automatizado es capaz de procesar miles de textos de manera simultánea (Ministerio de Asuntos Económicos y Transformación Digital et al., 2022). Diferentes estudios donde analizan la ciberagresión entre adolescente como por ejemplo el análisis de Álvarez-García et al. (2017) donde nos dice que la prevalencia de la ciber agresión entre los adolescentes y los diferentes de género tienen un impacto a largo plazo en la salud mental y bienestar digital.

La inteligencia artificial junto con una de sus capacidades la cual es el análisis de sentimientos ha llegado hacer su parte en alivio de los desafíos de lenguaje agresivo en redes sociales cuantificando la intensidad y relevancia de estos tonos emocionales en los textos logrando una mayor precisión que lo métodos tradicionales para detectar casos de acoso. De acuerdo con Bartolome (2021), la reducción del contenido dañino en las plataformas digitales seria de hasta un 62% , lo que pone de relieve su significativo impacto. No obstante, la detección de acoso de forma automatizada sigue siendo un desafío considerable, y esto se debe principalmente a la complejidad del lenguaje utilizado dentro las redes sociales. No tiene suficiente solo con analizar las palabras; el tono, el contexto y las referencias culturales también juegan un papel determinante. Varela Campos (2024) señala que cada publicación representa un pequeño universo comunicativo que muchas veces escapa a los enfoques clásicos de análisis. Además, fenómenos como el sarcasmo, la ironía y el lenguaje figurado complican aún más el trabajo de los sistemas automatizados, ya que un comentario aparentemente neutro puede esconder distintas capas de agresividad que solo un análisis más profundo logra revelar. Vinueza-Álvarez et al. (2023) nos dicen que aproximadamente el 73 % de los casos de ciberacoso involucran el uso de un lenguaje indirecto o altamente contextual, presentando un desafío importante para los sistemas automáticos de detección. Además, el ciberacoso no solo tiene un impacto directo en la salud mental de los adolescentes, sino que también promueve dinámicas de exclusión social entre ellos (Marín-Cortés, 2020).

Ante estos desafíos, el desarrollo de modelos de aprendizaje automático cada día evolucionan con el fin de abordar la complejidad creciente del lenguaje en entornos digitales. Si bien entonces, métodos clásicos nos muestran buenos resultados en tareas iniciales como son Naive Bayes o máquinas de vectores de soporte (SVM) estas mismas con el tiempo han evidenciado que sus capacidades son limitadas frente a los matices lingüísticos característicos de la comunicación en redes sociales. Por el contrario, enfoques más recientes como BERT (Bidirectional Encoder Representations from Transformers) nos han dado resultados positivos con un avance significativo en el procesamiento del lenguaje natural. A diferencia de modelos tradicionales que analizaban el texto de forma secuencial, BERT considera el contexto en ambas direcciones, lo que le permite interpretar mejor el significado completo de las oraciones. Esta característica ha demostrado ser especialmente valiosa en tareas como la clasificación de sentimientos y la detección de discursos de odio (Collarte Gonzalez, 2020). De igual manera, modelos generativos como GPT-2 y GPT-3 han mostrado una notable versatilidad para la clasificación de texto. Aunque inicialmente se diseñaron para generar texto, su capacidad para identificar patrones e intenciones en mensajes complejos también los ha hecho útiles en labores de clasificación, particularmente en la detección de amenazas o contenido agresivo (Chiu et al., 2022). En esta misma línea, investigaciones recientes han evidenciado que la combinación de BERT con modelos de tópicos mejora la clasificación de sentimientos en textos breves como los tweets, lo que refuerza su aplicabilidad en plataformas de redes sociales (Palani et al., 2021).Además, la efectividad de BERT en la detección de acoso en redes sociales ha sido probada en competencias internacionales de procesamiento del lenguaje natural, como SemEval-2017 (Das & Pedersen, 2024).

En investigaciones actuales como la de Amalia y Suyanto (2024) que explican el potencial de BERT en contexto similares, donde lograron un F1-score de 0,89 detectando lenguaje ofensivo y discurso de odio en redes sociales usando BERT, destacando su eficiencia en el manejo de textos con alto contenido emocional.

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

De manera similar, Amrenta-Segura et al. (2023) lograron obtener métricas significativas en tareas multimodales en datos textuales y visuales para identificar discurso de odio en memes. Asimismo, Aggarwal y Mahajan (2024) obtuvieron resultados superiores al 90% de precisión con un modelo conjunto entre BERT y SVM en ambientes sociales digitales reales.

Aunque existe una gran cantidad de investigaciones sobre la detección de lenguaje agresivo en redes sociales, muchos trabajos siguen enfocándose en modelos tradicionales y no terminan de resolver el problema del desbalance de clases en los datos. En este sentido, el presente estudio busca aportar algo distinto al comparar de manera práctica un modelo clásico de clasificación con uno de aprendizaje profundo como BERT, incorporando además la técnica SMOTE para equilibrar las clases. Con este enfoque se pretende no solo evaluar el rendimiento bajo escenarios más realistas, sino también analizar con más detalle los errores de clasificación, lo que podría servir como base para futuras estrategias de moderación automática de contenido.

A partir de esta problemática, se plantean las siguientes preguntas de investigación que orientan el presente estudio: ¿Qué tan efectivos son los modelos avanzados de inteligencia artificial, como BERT, en la detección automática de mensajes agresivos y amenazas en redes sociales?, ¿Cuál es el rendimiento de los modelos tradicionales (como Regresión Logística) en comparación con modelos avanzados en la clasificación de textos agresivos? y ¿Qué técnicas de balanceo de datos pueden mejorar la efectividad de los modelos en conjunto de datos desbalanceados con mensajes mayoritariamente no agresivos?

Como objetivo principal de la presente investigación es comparar de manera precisa y analítica el desempeño de los modelos tradicionales como lo es la regresión logística y los modernos como tenemos BERT para la clasificación y análisis de texto logrando detectar acoso, insultos y amenazas en línea. Para ello se utilizarán métricas de evaluación que son la precisión, la sensibilidad y la medida F1, estas métricas son de gran ayuda por que nos permiten valorar la capacidad de los modelos para clasificar correctamente distintos tipos de contenido textual. Se utilizará el conjunto de datos llamado Cyberbullying Dataset, disponible en la plataforma Kaggle, el cual contiene una amplia variedad de textos etiquetados, que abarcan desde mensajes con insultos y lenguaje ofensivo hasta contenidos parciales. Este conjunto de datos, ampliamente reconocido en la comunidad de procesamiento de lenguaje natural, permitirá entrenar y validar los modelos, evaluando su capacidad de generalización frente a textos con características diversas. Además, este enfoque permitirá responder las preguntas de investigación planteadas, proporcionando un marco práctico para comprender y abordar las limitaciones y capacidades de los modelos en la detección de textos agresivos.

Este estudio se apoya en la metodología CRISP, la cual estructura el proceso de análisis en seis etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Esta metodología facilita una organización clara del trabajo y también promueve la obtención de resultados coherentes y replicables. Esta investigación también busca aportar al desarrollo de modelos más eficaces para la detección automática de acoso y amenazas en entornos digitales, a la vez que se propone un marco práctico que pueda ser adoptado por plataformas tecnológicas para implementar estas soluciones de forma eficiente. En última instancia, este trabajo busca contribuir a la creación de espacios digitales más seguros, inclusivos y sostenibles para toda la comunidad de usuarios.

2. Materiales y Métodos

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) utilizada en este estudio es ampliamente aceptado en el campo del análisis de datos ya que nos ofrece marco estructurado pero lo suficientemente flexible como para adaptarse a diferentes tipos de proyectos, guiando cada etapa del proceso desde la definición del problema hasta la implementación de las soluciones. El proceso se divide en seis principales fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación; estas fases están diseñadas para asegurar la solidez del análisis y favorecer la replicabilidad de los resultados obtenidos. En la Figura 1 se detallan las fases que conforman la metodología CRISP-DM, la cual sirvió como base para organizar y desarrollar el proceso de minería de datos en este estudio. En esta sección, se describen las primeras cuatro etapas con un enfoque en los materiales, herramientas y métodos utilizados.

Figura 1. Etapas de la metodología CRISP-DM.
Fuente: Autoría propia basada en Rueda (2019)

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

Comprensión del negocio

El ciberacoso representa uno de los principales desafíos sociales en el entorno digital actual. A medida que el uso de las redes sociales crece exponencialmente, también lo hace el riesgo de conductas agresivas y amenazantes que afectan el bienestar psicológico y emocional de los usuarios. Este fenómeno es especialmente preocupante debido a su carácter persistente, su alcance global y la capacidad de los perpetradores para actuar de manera anónima. Por ello de manera urgente resulta fundamental avanzar con el desarrollo de herramientas automatizadas capaces de identificar de forma temprana y precisa contenidos que puedan representar un riesgo para los usuarios. En La Figura 2 se puede observar el porcentaje de personas que han reportado haber sido víctimas de ciberacoso en diversas plataformas de redes sociales según nos indica Panda Security, lo que pone de manifiesto la magnitud del problema en los entornos digitales. Los datos reflejan, además, que las plataformas con mayor nivel de interacción social y un fuerte componente visual tienden a ser más propensas a este tipo de conductas agresivas. La exposición al contenido digital sin regulación eficiente ha permitido que prácticas de ciberacoso continúen afectando a millones de usuario, lo que resalta la urgencia de soluciones tecnológicas efectivas (Panda Security, 2023).

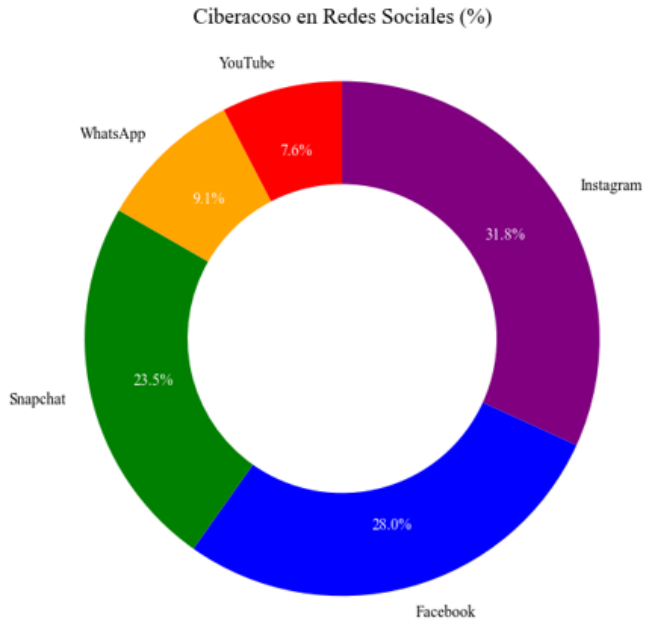

Figura 2. Porcentaje de usuarios que han experimentado ciberacoso en diferentes plataformas de redes sociales.
Fuente: Autoría propia basada en Panda Security (2023)

Comprensión de los datos

Para este estudio, se empleó el Cyberbullying Dataset, un conjunto de datos disponible en la plataforma Kaggle (https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset) y fue desarrollado por Saurabh Shahane, se ha seleccionado este conjunto de datos por que ha sido ampliamente utilizado en investigaciones previas, ya que contiene ejemplos reales de interacciones extraídas de redes sociales, organizados en archivos de formato CSV. Cada registro incluye, entre otros campos, el mensaje escrito por el usuario (etiquetado como "Text") y una variable llamada "oh_label", que clasifica el contenido como agresivo (valor "1") o no agresivo (valor "0"). Para ofrecer una visión más clara de su estructura, la Figura 3 presenta una muestra de las primeras filas del conjunto de datos.

	index	id \
0	5.74948705591165E+017	5.74948705591165E+017
1	5.71917888690393E+017	5.71917888690393E+017
2	3.90255841338601E+017	3.90255841338601E+017
3	5.68208850655916E+017	5.68208850655916E+017
4	5.75596338802373E+017	5.75596338802373E+017

	Text	Annotation	oh_label
0	@halalflaws @biebervalue @greenlinerzjm I read...	none	0.0
1	@ShreyaBafna3 Now you idiots claim that people...	none	0.0
2	RT @Mooseoftorment Call me sexist, but when I ...	sexism	1.0
3	@g0ssipsquirrelx Wrong, ISIS follows the examp...	racism	1.0
4	#mkr No No No No No No	none	0.0

Figura 3. Primeras filas del archivo toxicity_parsed_dataset.csv.
Fuente: Los autores.

Gracias a su diseño este conjunto de datos se convierte en una herramienta valiosa para tareas de clasificación de texto, ya que refleja una amplia variedad de actitudes y comportamientos observados en las redes sociales, desde interacciones ofensivas hasta conversaciones neutrales. En la Tabla 1 se detallan las principales características del Cyberbullying Dataset, incluyendo las categorías de mensajes presentes y su estructura original

Tabla 1. Características del Cyberbullying Dataset
Fuente: Los autores.

Elemento	Descripción
Conjunto de datos	Cyberbullying Dataset - Kaggle
Categorías del conjunto de datos	Ciberacoso directo, lenguaje ofensivo, neutro
Tamaño del conjunto de datos	Más de 100,000 registros
Formato de datos	CSV (valores separados por comas)

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

Preparación de los datos

A pesar de que el conjunto de datos ya está bien estructurado, los modelos necesitan de una etapa llamada preprocesamiento del texto, esta etapa nos asegura la calidad y coherencia de los datos utilizados en el análisis. El preprocesamiento de texto consiste en diversas tareas con el fin de mejorar la representatividad de los datos, en esta investigación la primera de esas tareas fue realizar una limpieza del texto ya que no aportaban nada al modelado como, por ejemplo, caracteres especiales, números y secuencias repetitivas, como comillas dobles y guiones bajos. Además, se convirtió todo el contenido textual a minúsculas, lo que ayudó a estandarizar su formato y evitar posibles inconsistencias durante el análisis. Luego se aplicó la tokenización, que consiste en segmentar el texto en palabras individuales facilitando el análisis de cada termino de manera aislada, otra tarea del preprocesamiento de texto fue eliminar las palabras vacías que son términos comunes como "the", "is" o "of", que no aportan valor significativo para la clasificación. Posteriormente se aplicó la lematización reduciendo las palabras a su forma base, unificando aquellas que comparten la misma raíz pero que presentan variaciones gramaticales. Este proceso de normalización contribuyó a un mejor desempeño del modelo al reducir el ruido en los datos.

Por último, el conjunto de datos presentaba un desbalance al tener mayor cantidad de mensajes no agresivos que de agresivos, para solucionar este problema y equilibrar el conjunto de entrenamiento se aplicó la técnica SMOTE la cual consiste en generar ejemplos sintéticos de la clase minoritaria mejorando la capacidad del modelo para identificar mensajes agresivos. Gracias al balanceo aplicado, se logró un aumento en las métricas de precisión y sensibilidad en la detección de lenguaje agresivo. En la Figura 4 se ilustra de manera esquemática todo el proceso de preprocesamiento llevado a cabo en este trabajo.

computacional. Para adaptar los textos al modelo fue necesario convertirlos en vectores numéricos, algo que se concretó a través de la técnica TF-IDF (Term Frequency-Inverse Document Frequency), característica que permite resaltar las palabras de más peso dentro de los documentos.

Para entender mejor cómo funciona la regresión logística, tenemos la representación gráfica y su función sigmoidea en la figura 5; función que resulta clave, pues permite transformar cualquiera de las entradas lineales en una probabilidad que se encuentra entre 0 y 1, facilitándose así la toma de decisiones en cuanto a la clasificación.

De esta manera comprenderemos mejor cómo funciona el algoritmo, y por qué a pesar de ser un modelo sencillo sigue siento útil para resolver problemas binarios a día de hoy, incluso ante modelos más complejos.

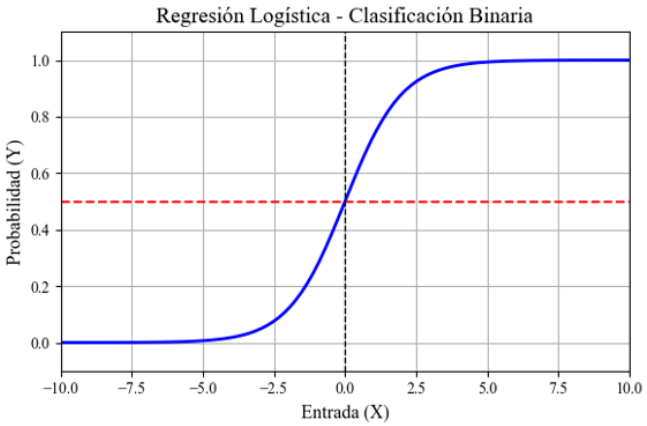

Figura 5. Función sigmoidea y grafica de la regresión logística.
Fuente: Los autores, adaptado de Logistic Function (s. f.)

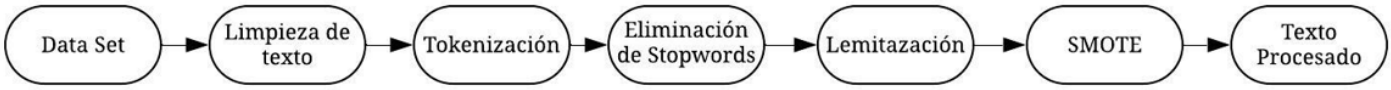

Figura 4. Etapas del preprocesamiento de texto.
Fuente: Los autores.

Modelado

Para el modelado comparamos modelos de clasificación de texto para la identificación de mensajes agresivos en línea, implementado el modelo tradicional conocido como Regresión Logística y modelos avanzados basados en aprendizaje profundo conocido como BERT.

Modelos tradicionales

Como primera instancia, este trabajo seleccionó técnicas tradicionales de machine learning ya que en este caso se pretendía establecer un marco de comparación sencillo pero eficaz. Se utilizó la regresión logística, un modelo ampliamente utilizado desde hace mucho tiempo debido a su simplicidad y bajo coste

Modelos avanzados

Aparte de las estrategias tradicionales, se optó por implementar el modelo BERT (Bidirectional Encoder Representations from Transformers) con los parámetros correspondientes de la tarea de clasificación binaria de la biblioteca Transformers de Hugging Face. El procedimiento realizado incluye como primera instancia la tokenización, que consiste en convertir los textos en tokens que el modelo es capaz de procesar. Luego se aplicó el entrenamiento que es un proceso de ajuste del modelo con los conjuntos de datos ya preprocesados eligiendo la función de pérdida correspondiente para la clasificación binaria, por último, se aplicó la optimización; para ellos se usó del optimizador AdamW y programadores de tasa de aprendizaje; con el objetivo de que el modelo sea lo

e-ISNN 2550-6730 | **Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025** | **DOI: 10.33936/isrtic.v9i1.7470**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

más eficiente posible. La Figura 6 muestra el rendimiento de los diferentes modelos en la detección de lenguaje ofensivo, observando las diferencias en precisión, sensibilidad y medida F1 entre el enfoque tradicional y el enfoque avanzado.

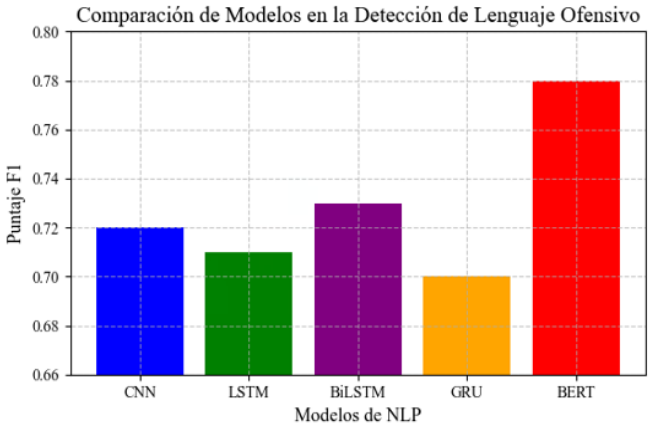

Figura 6. Comparación de modelos en la detección de lenguaje ofensivo.
Fuente: Los autores, basado en Sapora et al. (2019)

Detalles de implementación de modelos

En cuanto al modelo mas avanzado BERT, se utilizó la arquitectura bert-base-uncased que se adaptó a una tarea de clasificación binaria mediante la biblioteca Transformers de Hugging Face, para el proceso de tokenización se configuro una longitud máxima de 128 tokens con relleno automático y mecanismos de truncamiento.

El entramiento fue mediante 5 épocas, utilizando un tamaño de lote de 16 ejemplos y el optimizador AdamW con una tasa de aprendizaje de 2e-5, para la función de perdida se utilizó la entropía cruzada (CrossEntropyLoss).Se dividió los datos mediante en 80% para entrenar y 20% para validación, teniendo en cuenta que se mantengan el equilibrio de clases en ambos subconjuntos, en lugar de validación cruzada se utilizó una partición fija conocida con una semilla aleatoria (random_state=42) de esta manera se aseguro la reproducibilidad de los resultados. Para el balanceo de clases, se aplico varias estrategias en función de modelo. Para la regresión logística, se utilizo SMOTE sobre los vectores TF-IDF previamente extraídos, donde se generó ejemplos sintéticos de la clase minoritaria para equilibrar el conjunto de entramiento, al contrario del modelo avanzado se escogió un método de sobremuestreo aleatorio, duplicando las instancias de la clase minoritaria directamente sobre el conjunto original antes del proceso de tokenización.

Esta decisión permitió mantener la integridad semántica de los textos y mejorar el balance de clases sin introducir datos sintéticos en la arquitectura basada en aprendizaje profundo.

Ética y Consideraciones de Privacidad

En la investigación se adoptó medidas necesarias para garantizar la privacidad y confidencialidad de los datos debido a que el conjunto de datos tendría información sensible; todos los textos procedieron a ser anonimizados y los resultados se utilizaron únicamente con el fin investigativo respetando la normativa ética correspondiente.

3. Resultados y Discusión

Para evaluar qué tan bien están funcionando los modelos, observamos cuatro indicadores de rendimiento principales: sensibilidad (recall), medida F1 y exactitud, las demás mediciones se determinaron para cada clase utilizando promedios macro y ponderados, con la evaluación más justa que tiene en cuenta el desbalance inicial en los datos. La función de clasificación_ report de la biblioteca de Scikit-Learn le brinda estas métricas directamente de las etiquetas reales y las predicciones del modelo. Para Bert, la clase más probable se eligió mediante la que tiene mayor oportunidad (Argmax), y no había necesidad de establecer ningún límite manual. El estudio también se mejoró al incluir una matriz de confusión y gráficos que muestran el rendimiento para cada categoría, lo que facilita ver dónde los modelos tuvieron éxito o fallaron.

3.1. Desempeño del modelo BERT

El modelo BERT (Bidirectional Encoder Representations from Transformers), entrenado sobre un conjunto de datos balanceado mediante la técnica SMOTE, evidenció un rendimiento robusto en la clasificación binaria de mensajes agresivos y no agresivos. Para el entrenamiento se utilizaron cinco épocas y un tamaño de lote de 16, lo cual permitió obtener resultados estables sin indicios de sobreajuste. Durante el proceso de entrenamiento, se registró una disminución progresiva de la función de pérdida, comenzando en 0.41 y reduciéndose hasta 0.03 en la quinta época. Este comportamiento sugiere que el modelo logró aprender los patrones relevantes del lenguaje sin presentar problemas de oscilación o estancamiento. La Figura 7 ilustra la curva de pérdida obtenida, confirmando la correcta convergencia del modelo.

El modelo avanzado utilizado que es el BERT nos dio una respuesta muy buna con una precisión general del 93% y una medida-F1 promedio del 0.93, manteniendo un equilibro en ambas clases. En la Tabla 2 podemos ver las métricas evaluadas, obteniéndose una precisión mayor para la clase no agresiva

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

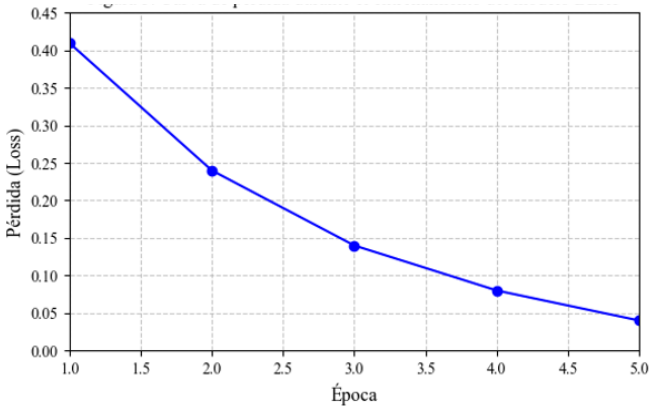

Figura 7. Curva de pérdida durante el entrenamiento del modelo BERT.
Fuente: Los autores.

(0.96) y una mayor sensibilidad para la clase agresiva (0.96), lo que indica que se ha alcanzado un equilibrio entre la detección y la precisión.

Tabla 2. Métricas de evaluación del modelo BERT.
Fuente: Los autores.

Clase	Precisión	Sensibilidad	Medida F1	Soporte
0 (No agresivo)	0.96	0.90	0.93	2301
1 (Agresivo)	0.90	0.96	0.93	2300
Promedio macro	0.93	0.93	0.93	
Promedio ponderado	0.93	0.93	0.93	
Precisión global	—	—	0.93	4601

Los resultados obtenidos permiten concluir que BERT es un modelo efectivo para la detección automatizada de ciberacoso, especialmente en entornos textuales complejos como las redes sociales, donde la ambigüedad semántica y la variabilidad lingüística representan desafíos significativos para los clasificadores tradicionales.

3.2. Evaluación por clase y métricas generales

El modelo BERT se comportó de una manera equilibrada cuando se pretendía clasificar si los mensajes eran agresivos o no agresivos, de modo que la precisión global que logró alcanzar fue del 93% y obtuvo una medida F1 promedio del 0.93. Esas medidas indican que se trató de un modelo fiable y robusto, es decir, uno que supo generalizar bien sobre el lenguaje variable de las redes sociales.

La distribución de aciertos y errores que se puede observar se puede resumir en una matriz de confusión que se puede ver en la Figura 8 donde se pueden apreciar que el modelo supo identificar bien 2238 mensajes agresivos, así como 2022 mensajes no agresivos; no obstante, también suma 279 falsos positivos es decir que ha clasificado como agresivos a mensajes neutros, así como 62 falsos negativos que no han sabido detectar que contenía agresiones reales en los mensajes.

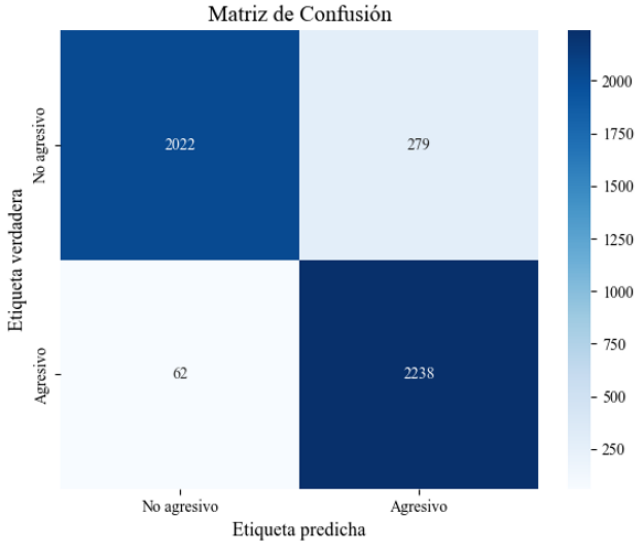

Figura 8. Matriz de confusión del modelo BERT.
Fuente: Los autores.

Las métricas por clase también fueron representadas gráficamente en la Figura 9, destacándose una mayor precisión en la clase no agresiva (0.96) y una mayor sensibilidad en la clase agresiva (0.96). Esta combinación sugiere que el modelo es particularmente efectivo para la detección de lenguaje ofensivo, sin comprometer su capacidad de discernir entre lo agresivo y lo neutro.

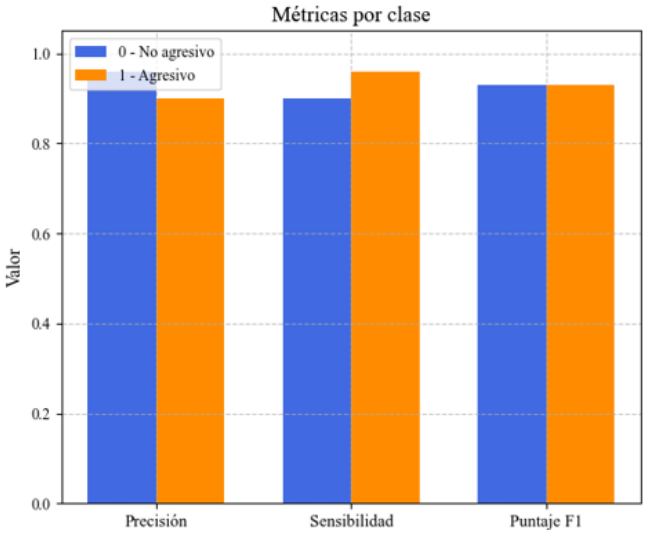

Figura 9. Métricas por clase del modelo BERT: precisión, sensibilidad y medida F1.
Fuente: Los autores.

El adecuado rendimiento equilibrado del modelo en ambas clases es un elemento primordial para aplicaciones de detección

e-ISNN 2550-6730 | **Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025** | **DOI: 10.33936/isrtic.v9i1.7470**

automática de ciberacoso, garantizando así una baja frecuencia de errores sobre un entorno sensible como son los espacios digitales usados por adolescentes y jóvenes.

3.3. Comparación entre modelos tradicionales y avanzados

Con la intención de valorar cómo se comportan las distintas técnicas para la detección automática del lenguaje agresivo utilizado en redes sociales, se realizó una comparativa entre dos enfoques diferentes, uno el modelo tradicional de Regresión Logística con vectores TF-IDF y, por otro lado, el modelo avanzado BERT basado en redes neuronales profundas. Ambos modelos se entrenaron y evaluaron sobre el mismo conjunto de datos que fue previamente balanceado usando SMOTE.

En la Tabla 3 se muestran los resultados de ambas estrategias donde se describen las métricas de precisión, sensibilidad, medida F1 y exactitud general. Mientras que la Regresión Logística obtuvo una medida F1 promedio de 0.83, el modelo BERT supera ese rendimiento con una media de 0.93, lo que pone en evidencia una mejora sustancial en la detección.

Tal como se muestra en la Figura 10, la matriz de confusión del modelo de Regresión Logística presenta una cantidad considerable de errores de clasificación de mensajes agresivos (241 falsos negativos y 351 falsos positivos). Este comportamiento se ve reflejado igualmente en la Figura 11, donde se aprecia una gran diferencia entre las métricas de ambas clases, sobre todo en la

Tabla 3. Comparación del rendimiento entre modelos tradicionales y avanzados.
Fuente: Los autores.

Modelo	Precisión	Sensibilidad	Medida F1	Exactitud
Regresión Logística (TF-IDF)	0.83	0.82	0.83	0.82
BERT	0.93	0.93	0.93	0.93

precisión de la clase 1 (agresivo), que apenas logra superar el 0.70.

Al contrario, el modelo BERT presenta mayores precisiones y sensibilidades para ambas clases, además de que equilibra el rendimiento entre ambas, reduciendo el desbalance de rendimiento entre los mensajes clasificados como agresivos y no agresivos. Esta mejora de rendimiento se puede aventurar que es debida a la capacidad contextual de BERT, que mantiene las relaciones de las palabras, sin depender únicamente de su frecuencia, como lo hace el TF-IDF.

Como complemento de los resultados por clase, la Figura 12 permite ver de forma global el rendimiento de los modelos. El modelo BERT muestra un rendimiento superior al modelo de regresión logística para todas las métricas evaluadas, mostrando su superioridad para tareas de clasificación en lenguaje humano, donde el contexto semántico y la estructura gramatical son determinantes para la interpretación del mensaje.

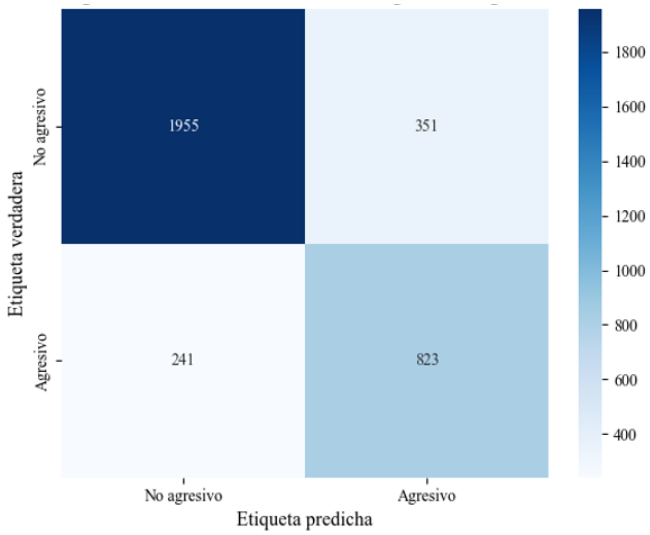

Figura 10. Matriz de confusión – Regresión logística.
Fuente: Los autores.

Por lo que podemos concluir que el modelo basado en BERT es una solución mejor para tareas de detección automática de ciberacoso, sobre todo en el caso de textos donde el lenguaje es ambiguo, con muchas expresiones coloquiales y sarcasmos, presente en los mensajes que nos ofrecen las redes sociales.

3.5. Discusión

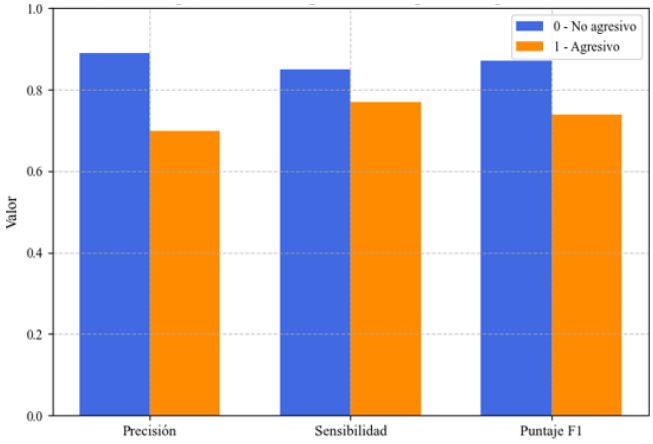

Figura 11. Matriz por clase – Regresión logística.
Fuente: Los autores.

Las comprobaciones comparadas entre los dos modelos llevados a cabo a través de experimentos de evaluación en el rendimiento de cada uno de ellos permitieron identificar diferencias considerables, no solo en su rendimiento global, sino en la manera en que cada modelo lleva a cabo la tarea implicando mensajes diferentes. Aunque ambos modelos aportaron resultados considerados como aceptables, resultó un hecho el que BERT mostró una mayor capacidad para detectar lenguaje agresivo, especialmente en aquellos ejemplos que requerían un contexto y

e-ISNN 2550-6730 | **Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025** | **DOI: 10.33936/isrtic.v9i1.7470**

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

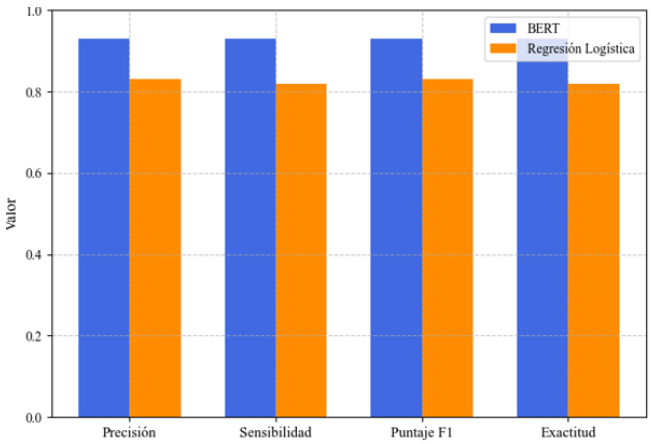

Figura 12. Comparación de métricas globales entre BERT y Regresión logística.
Fuente: Los autores.

matices semánticos elevados. En este sentido, la arquitectura de atención bidireccional, ayudo a analizar de mejor manera cuando se presentan textos ambiguos o expresiones no literales. Este tipo de comprensión debe ser considerarse en tareas tales como la moderación de contenido en las redes sociales, donde entender el trasfondo del mensaje puede marcar la diferencia.

Por otro lado, la regresión logística que se basa solo en representaciones TF-IDF, mostró carencias en el entendimiento de los mensajes con textos no literales. La regresión logística podemos decir que alcanzo una precisión razonable en general, pero cometió más errores al calcificar mensajes agresivos como no agresivos, y viceversa. Estos errores podemos observar en la matriz de confusión, en la que la métrica de sensibilidad era considerablemente menor para mensajes agresiones.

Aun así, gracias a su simplicidad y al bajo coste computacional, la regresión logística se puede seguir considerando válida en ambientes donde el enfoque rápido y eficiente por encima de la complejidad del modelo. Otro de los factores que afectan al rendimiento de ambos enfoques es la técnica SMOTE al balancear las clases. En un primer momento el conjunto de datos aporta un gran desbalance, lo que dificultaba que los modelos aprendieran patrones representativos de la clase minoritaria. Luego de aplicar el balanceo se observaron mejoras en métricas como la sensibilidad y la precisión. Este resultado pone de relieve la necesidad de poner en marcha estrategias de equilibrio, muy especialmente en tareas en las que las clases no están balanceadas.

No obstante, y aunque se haya avanzado mucho, también se han podido identificar limitaciones que no conviene perder de vista. Tanto BERT como la técnica de regresión logística tienen ciertos problemas a la hora de poner a prueba mensajes que contengan sarcasmo o del lenguaje implícito. No son de hecho los casos con mayor frecuencia. Sin embargo, sus efectos sobre la precisión han podido deducirse. Esta observación muestra que en futuras investigaciones podría ser beneficiosa la integración de análisis pragmáticos o de componentes que tengan en cuenta aspectos socioculturales del lenguaje digital. En este sentido, la visualización de los resultados en forma de gráficos y matrices no sólo facilita interpretarlas en términos cuantitativos, sino que también sirve para entender de manera más perfunda cómo y por qué fallan los modelos en determinados casos.

En comparación de estos resultados y de sus respectivas limitaciones con trabajos anteriores, se muestra una tendencia a la hora de usar modelos avanzados que se mantiene presente. Zampieri et al. (2019) presentan datos en los que BERT se asienta de forma clara como un modelo que, por sí mismo, supera de manera categórica los métodos clásicos en la precisa y la confiabilidad del discurso de odio, resultados que coinciden con los hallados en este trabajo. Por su parte, Pamungkas et al. (2020) también reportan que BERT, una vez más, obtiene tasas mucho más frecuentes y precisas de los modelos clásicos SVM y Naive Bayes. Estos precedentes nos remiten a la premisa inicial de que las arquitecturas de Transformers como BERT aportan ventajas reales para la captura de matices emocionales y contextuales en las redes sociales, llegando a ser una alternativa consolidada sobre modelos clásicos.

Si se comparan estos resultados con trabajos recientes que también usaron modelos BERT, se puede ver que el rendimiento logrado en este estudio se mantiene dentro del rango alto. Por ejemplo, Amalia y Suyanto (2024) lograron un F1-score de 0,89 al detectar lenguaje ofensivo, mientras que el presente estudio alcanzó un 0,93 trabajando con textos agresivos más variados. A diferencia de otras investigaciones, aquí se usó la técnica de sobre muestreo (SMOTE) en el modelo clásico y un enfoque práctico de duplicación en BERT, lo cual permitió manejar mejor el desbalance. Además, se estructuró todo el proceso bajo la metodología CRISP-DM, que no siempre se ve aplicada en este tipo de estudios. Aunque no se propone un nuevo modelo, la comparación práctica entre un enfoque tradicional y uno moderno con técnicas complementarias ofrece una base útil para trabajos similares en el futuro.

4. Conclusiones

Comparar la regresión logística y BERT para la detección automática de lenguaje agresivo en redes sociales ayudó a entender mejor la capacidad y límites de cada modelo. A lo largo del análisis, quedó bastante claro que BERT lleva ventaja en la mayoría de las métricas, sobre todo cuando se trata de captar el contexto o la intención detrás de un mensaje. Se percibió, en particular, en la detección de mensajes agresivos, donde logró

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

mantener un mejor equilibrio entre falsos positivos y falsos negativos, al contrario del modelo tradicional donde tuvo una respuesta más baja. Parece que la arquitectura basada en atención bidireccional le da a BERT una ventaja para captar matices que, de otro modo, podrían pasar desapercibidos.

En cambio, la regresión logística, pese a su simplicidad, demostró ser una opción razonable en escenarios donde los recursos computacionales no son los mejores. Aunque tuvo un desempeño aceptable en tareas más básicas, tuvo una respuesta baja cuando el análisis semántico debía ser más profundo. En este punto, conviene recordar que no todo el éxito de los modelos se debe a su estructura interna: aplicar SMOTE para balancear las clases también fue decisivo. Sin ese paso, probablemente los resultados habrían sido más modestos, especialmente para la clase minoritaria.

Más allá de los números, este trabajo también aporta una mirada práctica sobre cómo estas herramientas enfrentan un problema tan delicado como el ciberacoso.

Y no todo depende del modelo elegido: la limpieza de los datos, la forma en que se etiquetan los mensajes y hasta las particularidades del lenguaje que se usa en cada red social terminan pesando bastante en el resultado final. Aunque los modelos rindieron bien en general, hay temas pendientes que no se pueden ignorar, como el sarcasmo, el doble sentido o las expresiones cargadas de referencias culturales. Pensando en trabajos futuros, se podría pensar en la opción de incorporar modelos híbridos o estrategias que permitan entender mejor otros casos más complejos. Al final del día, entender la comunicación humana y toda su ambigüedad sigue siendo uno de los mayores desafíos para la inteligencia artificial.

El siguiente estudio más allá de comparar métricas también da un aporte, una guía practica sobre como realizar un proceso completo de clasificación de texto en problemas actuales como el ciberacoso. Integrar CRISP-DM ayudo a que se mantenga una clara secuencia desde la compresión del problema hasta finalmente la evaluación, lo que mejora poder replicarlo. Además, usar técnicas de balanceo como SMOTE y balanceo directo como se uso en BERT les da la facilidad a otros autores el poder adaptar este tipo de proyectos según sus necesidades. A pesar de que el modelo no es nuevo la forma de comparar y ajustar ambos enfoques puede ser útil para trabajos futuros donde se requiera precisión en la moderación de contenido.

Contribución de los autores

Kevin Alexander Mendoza Campoverde: Conceptualización, Metodología, Curación de datos, Software, Investigación, Visualización, Redacción – borrador original del artículo. **Javier Valentín Hurtado González**: Conceptualización, Metodología, Redacción – borrador original del artículo, Análisis formal. **Rodrigo Fernando Morocho Román**: Administración del proyecto, Supervisión, Revisión y edición del artículo. **Wilmer Braulio Rivas Asanza**: Validación – Verificación y Visualización.

Conflictos de interés

Los autores declaran no tener ningún conflicto de interés.

Anexos

Anexo A. Evaluación adicional del modelo BERT

A.1. Ejemplos de textos mal clasificados

Ejemplos mal clasificados:

	Text	Real	Prediccion
2	RT @ThatSabineGirl: These abusers on 8chan hav...	0	1
11	RT @Rildom1: @YesYoureSexist also get your shi...	1	0
19	@SevilZadeh Here is a great example of Mohamme...	0	1
61	I'm not a Misandrist but Males in General can ...	0	1
67	@lilrabmike @Chickowits said so bravely from t...	0	1
80	@JamesBolton11 Yup.	1	0
91	@Caspasia_Boleyn and what a judgmental ass you...	0	1
98	#YesAllWomen see fire when men make excuses fo...	0	1
102	@Ammaawah @jm111t More sexism. Keep women "cha...	0	1
104	RT @ESMART234: ISIS: Jihadist police beat up w...	0	1

Anexo B. Preparación del conjunto de datos y balanceo de clases

B.1. Distribución original del conjunto de datos y distribución después de aplicar el balanceo

```
Distribución original:
oh_label
0.0    11501
1.0     5347
Name: count, dtype: int64

Distribución después del balanceo:
oh_label
1.0    11501
0.0    11501
Name: count, dtype: int64
```

B.2. Fragmento de código aplicado para el balanceo

```
# Balancear dataset con sobremuestreo de la clase minoritaria
minority_df = df[df["oh_label"] == 1]
majority_df = df[df["oh_label"] == 0]
minority_upsampled = minority_df.sample(len(majority_df), replace=True, random_state=42)
df_balanced = pd.concat([majority_df, minority_upsampled]).sample(frac=1, random_state=42).reset_index(drop=True)

print("\nDistribución después del balanceo:")
print(df_balanced["oh_label"].value_counts())
```

Referencias bibliográficas

Aggarwal, P., & Mahajan, R. (2024). Shielding social media: BERT and SVM unite for cyberbullying detection and classification. *Journal of Information Systems and Informatics, 6*(2). https://doi.org/10.51519/journalisi.v6i2.692

Álvarez-García, D., Barreiro-Collazo, A., & Núñez, J.-C. (2017). Ciberagresión entre adolescentes: Prevalencia y diferencias de género. *Comunicar: Revista Científica de Comunicación y Educación, 25*(50), 89–97. https://doi.org/10.3916/C50-2017-08

Amalia, F. S., & Suyanto, Y. (2024). Offensive language and hate speech detection using BERT model. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 18*(4), Article 4. https://doi.org/10.22146/ijccs.9984191A.1

Armenta-Segura, J., Núñez-Prado, C. J., Sidorov, G. O., Gelbukh, A., & Román-Godínez, R. F. (2023). Ometeotl@

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470

INFORMÁTICA Y SISTEMAS
Revista de Tecnologías de la Informática y las Comunicaciones

Facultad de Ciencias Informáticas
Universidad Técnica de Manabí
Av. Urbina y Che Guevara, Portoviejo, Ecuador
revista.iys@utm.edu.ec

Multimodal Hate Speech Event Detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained BERT models over text. En Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (CONSTRAINT) (pp. 53–59). https://aclanthology.org/2023.case-1.7/

Bartolomé, M. (2021). Redes sociales, desinformación, cibersoberanía y vigilancia digital: Una visión desde la ciberseguridad. *RESI: Revista de Estudios en Seguridad Internacional, 7*(2), 167–185.

Chérrez, W. E. M., & Avila-Pesantez, D. F. (2021). Ciberseguridad en las redes sociales: Una revisión teórica. *Revista Uniandes Episteme, 8*(2), Article 2.

Collarte Gonzalez, I. (2020). Procesamiento del lenguaje natural con BERT: Análisis de sentimientos en tuits [Trabajo de Fin de Grado, Universidad Carlos III de Madrid]. https://e-archivo.uc3m.es/rest/api/core/bitstreams/a10e2295-b239-4305-aad1-1570259607bf/content

Das, R. K., & Pedersen, T. (2024). SemEval-2017 Task 4: Sentiment analysis in Twitter using BERT (No. arXiv:2401.07944). arXiv. https://doi.org/10.48550/arXiv.2401.07944

Logistic function. (s. f.). Scikit-Learn. Recuperado 12 de febrero de 2025, de https://scikit-learn/stable/auto_examples/linear_model/plot_logistic.html

Marín-Cortés, A. (2020). Las fuentes digitales de la vergüenza: Experiencias de ciberacoso entre adolescentes. *The Qualitative Report, 25*(1), 166–180. https://doi.org/10.46743/2160-3715/2020.4218

Ministerio de Asuntos Económicos y Transformación Digital, Red.es, & Observatorio Nacional de Tecnología y Sociedad. (2022). Beneficios y riesgos del uso de Internet y las redes sociales. Observatorio Nacional de Tecnología y Sociedad. https://doi.org/10.30923/094-22-017-3

ONTSI. (2022). Violencia digital de género: Una realidad invisible. https://www.ontsi.es/es/publicaciones/violencia-digital-de-genero-una-realidad-invisible-2022

Pamungkas, E., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management, 57*, 102360. https://doi.org/10.1016/j.ipm.2020.102360

Rueda, J. F. V. (2019, noviembre 4). CRISP-DM: Una metodología para minería de datos en salud. HealthDataMiner. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Sapora, S., Lazarescu, B., & Lolov, C. (2019). Absit invidia verbo: Comparing deep learning methods for offensive language (No. arXiv:1903.05929). arXiv. https://doi.org/10.48550/arXiv.1903.05929

Security, P. (2023, marzo 13). 52 estadísticas y datos alarmantes sobre el ciberacoso. Panda Security Mediacenter. https://www.pandasecurity.com/es/mediacenter/52-estadisticas-ciberacoso/

Varela Campos, E. (2024). Análisis de la privacidad y seguridad en las redes sociales en un mundo de ciberdelitos. https://repositorio.comillas.edu/xmlui/handle/11531/80324

Vinueza-Álvarez, C., Acosta-Uriguen, M. I., & Sigua, J. F. L. (2023). Análisis de clusterización en datos de encuestas sobre ciberacoso. *Revista Tecnológica - ESPOL, 35*(2), Article 2. https://doi.org/10.37815/rte.v35n2.1055

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. En J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 1415–1420). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1144

e-ISNN 2550-6730 | Vol. 9, Núm. 1 (82-92): Enero-Junio, 2025 | DOI: 10.33936/isrtic.v9i1.7470