http://doi.org/10.46480/esj.10.2.295

Evaluación de Redes Neuronales para la Identificación de Discurso de Odio en Español Utilizando Características Estilométricas

Evaluation of Neural Networks for the Identification of Hate Speech in Spanish Using Stylometric Features

León-Granizo Oscar Darío[1], Jaramillo Ramírez Valeria Nicol[1], Lamilla Vera Kevin Javier[1], Espin Riofrío Cesar[1], y Reyes Zambrano Gary[1]

RESUMEN

Contexto: Esta investigacion se centra en analizar la efectividad de las redes neuronales en la deteccion de discurso de odio en español, utilizando caracterısticas estilometricas y considerando la diversidad linguıstica de Ecuador. El objetivo principal fue evaluar y analizar la precision de diferentes modelos de redes neuronales en esta tarea especıfica de clasificacion de textos. **Metodo:** Se implementaron y compararon cuatro arquitecturas de redes neuronales: Red Neuronal Convolucional (CNN), Red Neuronal Recurrente (RNN), Memoria a Corto y Largo Plazo (LSTM) y Transformer, aplicando la metodologıa CRISP-DM para estructurar las tareas necesarias en cada fase y sus interrelaciones. **Resultados:** Los resultados revelaron diferencias significativas en el desempeño de los cuatro modelos. El modelo Transformer se destaca como la alternativa mas eficiente, logrando un puntaje F1 ponderado mas alto (0,88) y un AUC de 0,92, lo que indica un equilibrio optimo. Tanto CNN como RNN alcanzaron una precision global del 87 %, mientras que LSTM obtuvo la menor precision, con un 86 %. **Conclusiones:** En consecuencia, se recomienda la adopcion del modelo Transformer para la deteccion automatica de discurso de odio en textos en español, empleando caracterısticas estilometricas y considerando la diversidad linguıstica de Ecuador. Adicionalmente, se sugiere continuar evaluando el modelo en un entorno de produccion para identificar posibles sesgos y realizar los ajustes necesarios. La investigacion futura incluye incorporar textos en multiples idiomas y desarrollar programas educativos dirigidos a estudiantes, fomentando ası una mayor concienciacion sobre el discurso de odio y su posible impacto.

Palabras clave: Inteligencia artificial, Redes neuronales, Estilometrıa, Discurso de odio, Clasificacion de textos

ABSTRACT

Context: This research focuses on analyzing the effectiveness of neural networks in detecting hate speech in Spanish, using stylometric features and taking into account Ecuador's linguistic diversity. The main objective was to evaluate and analyze the accuracy of different neural network models in this specific text classification task. **Method:** Four neural network architectures were implemented and compared: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Transformer, applying the CRISP-DM methodology to structure the necessary tasks in each phase and their interrelationships. **Results:** The results revealed significant differences in the performance of the four models. The Transformer model stands out as the most efficient alternative, achieving a higher weighted F1 score (0.88) and an AUC of 0.92, indicating an optimal balance. Both CNN and RNN reached an overall accuracy of 87%, while the LSTM obtained the lowest accuracy at 86%. **Conclusions:** Consequently, the adoption of the Transformer model is recommended for automatic hate speech detection in Spanish texts, employing stylometric features and considering Ecuador's linguistic diversity. Furthermore, it is suggested to continue evaluating the model in a production environment to identify potential biases and make necessary adjustments. Future research includes incorporating texts in multiple languages and developing educational programs aimed at students, thereby fostering greater awareness of hate speech and its potential impact.

Keywords: Artificial intelligence, Neural networks, Stylometry, Hate speech, Text classification

Fecha de recepción: Mayo 13, 2026
Fecha de aceptación: Junio 30, 2026

[1]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Guayaquil, Ecuador. E-mails: oscar.leong@ug.edu.ec; valeria.jaramillor@ug.edu.ec; kevin.lamillav@ug.edu.ec; cesar.espinr@ug.edu.ec; gary.reyesz@ug.edu.ec

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence.

Cómo citar: León-Granizo, O. D., Jaramillo-Ramírez, V. N., Lamilla-Vera, K. J., Espín-Riofrío, C., & Reyes-Zambrano, G. (2026). Evaluación de Redes Neuronales para la Identificación de Discurso de Odio en Español Utilizando Características Estilométricas. *Ecuadorian Science Journal, 10(2), 1–10, Septiembre–2026.* DOI: https://doi.org/10.46480/esj.10.2.295

Introducción

A finales de la primera década de los 2000 e inicios de los 2010, las redes sociales como Facebook, Twitter y YouTube comenzaron a ganar popularidad global, convirtiéndose en espacios virtuales en los que personas de todas las edades crearon perfiles para compartir experiencias, historias personales y recuerdos con sus contactos (Fu et al., 2025). La llegada de los dispositivos móviles y la introducción de los teléfonos inteligentes facilitaron la creación y difusión instantánea de contenido, lo que llevó a un número cada vez mayor de usuarios a unirse a estos entornos digitales (Alghazzawi et al., 2025).

A medida que aumentó el número de usuarios activos, las plataformas de redes sociales enfrentaron el desafío de regular las interacciones en línea. Por un lado, comenzaron a desarrollar políticas de moderación de contenido orientadas a reducir el ciberacoso y el discurso de odio (Jiménez-Díaz and Del Rey, 2025); sin embargo, esto generó controversias relacionadas con la libertad de expresión. En el esfuerzo por no restringir excesivamente las opiniones, muchos comentarios ofensivos pasaron desapercibidos o no fueron sancionados adecuadamente. Así, las redes sociales se convirtieron en escenarios de confrontación y difusión de mensajes de odio basados en creencias políticas, culturales y religiosas (Assimakopoulos and Serafis, 2025; Perry and Olsson, 2009).

La comunicación violenta se entiende como cualquier forma de expresión en línea que pueda considerarse agresiva o violenta, esté o no motivada por razones discriminatorias (Miró Llinares, 2016; Esteve Bañón, 2020). Considerando que el discurso de odio abarca cualquier forma de comunicación que promueva o incite a la discriminación o la violencia contra un individuo o grupo por razones de raza, etnia, nacionalidad o religión, es importante reconocer que este fenómeno puede manifestarse de diversas maneras (Shohan et al., 2025). Puede expresarse de forma explícita, mediante amenazas o insultos directos, o de manera más sutil, a través de burlas o expresiones que normalizan la exclusión y el rechazo de grupos vulnerables (Hassan et al., 2025).

Sin embargo, moderar el contenido ofensivo plantea múltiples retos, ya que es esencial considerar el contexto y la forma en que dicho contenido es generado (Álvarez Benjumea, 2025). Es aquí donde la diversidad linguística de cada país entra en juego, influyendo en la manera en que se expresa el discurso ofensivo. En el caso de Ecuador, donde coexisten múltiples regiones con particularidades léxicas y sintácticas, resulta esencial incorporar características estilométricas que den cuenta de estas diferencias. Este enfoque permite identificar patrones en la escritura del discurso ofensivo con mayor precisión (Zapatier et al., 2024).

La revisión manual de los comentarios o discursos publicados por los usuarios en plataformas digitales, con el fin de determinar si contienen discurso de odio, resulta extremadamente desafiante para los moderadores (Ureña-López et al., 2024). El número de usuarios activos crece exponencialmente, lo que genera un gran volumen de interacciones y hace prácticamente imposible monitorear y clasificar todo el contenido de manera oportuna (Patel et al., 2025a).

Por otra parte, la falta de comprensión de la diversidad linguística y cultural de la región donde se genera el contenido dificulta la correcta interpretación de las expresiones y la detección de matices ofensivos (Park et al., 2024). Debido a estas desventajas, es probable que los moderadores de las plataformas tomen decisiones incorrectas, lo que podría generar desconfianza en los usuarios y, a su vez, facilitar la difusión de discurso de odio en el entorno digital.

Ante este contexto, surge la necesidad de explorar soluciones innovadoras que ayuden a los moderadores a detectar de forma automática si un comentario de un usuario puede considerarse discurso de odio, monitoreando de este modo su comportamiento dentro de la plataforma (Romero-Vega et al., 2021). En este sentido, las tecnologías de inteligencia artificial, en particular las redes neuronales, representan una alternativa prometedora para automatizar y mejorar la detección de discurso ofensivo. Estas herramientas tienen un gran potencial para procesar grandes volúmenes de texto, identificar patrones y evaluar con precisión cuándo un comentario puede clasificarse como ofensivo (Suntaxi et al., 2024).

Este estudio se enfoca en evaluar y analizar la efectividad de diferentes arquitecturas de redes neuronales en la detección de discurso de odio en textos en español, utilizando características estilométricas y considerando la diversidad linguística de Ecuador. Específicamente, se implementarán y compararán cuatro modelos de redes neuronales: Redes Neuronales Convolucionales (CNN), Redes Neuronales Recurrentes (RNN), Memoria a Corto y Largo Plazo (LSTM) y Transformer. Cada uno de estos modelos posee características únicas que los hacen potencialmente adecuados para la tarea de identificar patrones secuenciales en las características estilométricas de los textos, y su comparación permitirá determinar cuál arquitectura es la más eficaz para alcanzar este propósito (Patel et al., 2025b). La investigación emplea la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para estructurar las fases del proceso de análisis de datos. Esta metodología proporciona un enfoque sistemático y riguroso para el desarrollo de modelos de aprendizaje automático, asegurando que cada etapa del proyecto, desde la comprensión del negocio hasta la implementación, sea abordada de manera integral y metódica (Gonçalves et al., 2025).

Materiales y Métodos

Antecedentes del estudio

En la tesis de maestría titulada *"Modelo de detección de discurso de odio en Ecuador mediante clasificación supervisada de tweets y técnicas de NLP"* (Suntaxi Recalde, 2022), se propone desarrollar un modelo de detección de discurso de odio (HS) mediante la clasificación supervisada de tweets, aprovechando recursos avanzados de procesamiento del lenguaje natural (NLP). El objetivo es identificar contenido de odio en tweets generados en Ecuador, evaluando el esquema de Word Embedding más adecuado para la representación semántica de este tipo de textos en la región y seleccionando el modelo de clasificación supervisada con mejor rendimiento. Los tweets fueron recolectados a través de la API de Twitter usando el método de extracción en tiempo real. Los resultados muestran que BERT, con aprendizaje por transferencia, fue el modelo más efectivo para clasificar tweets de odio en el contexto ecuatoriano, alcanzando una precisión de 0,88 y un alto desempeño en métricas de recall. En el análisis comparativo, Word2Vec utilizó el módulo gensim para generar representaciones vectoriales y fue evaluado con varios modelos de clasificación supervisada como XGBoost (alcanzando una precisión de 0,71), Random Forest y SVM. Aunque XGBoost mostró buenos resultados de precisión, algunos modelos tuvieron dificultades para identificar casos específicos de odio.

En la tesis de maestría titulada *"Detección de Discurso de Odio Contra la Comunidad LGTBI+ Mexicana Mediante el Uso de Arquitecturas Transformer"* (Fernández Rosauro, 2023), se busca contribuir a la igualdad para las personas LGTBI+ mediante la detección de contenido LGTBI-fóbico en redes sociales. El objetivo del proyecto es explorar el rendimiento de los modelos Transformer en la detección de mensajes de odio, comparándolos con modelos de clasificación clásicos. Para ello, se obtuvieron dos conjuntos de datos de tweets en español mexicano de la competencia HOMO-MEX 2023. Los resultados de esta tesis muestran que los modelos Transformer específicos del español, como RoBERTuito y BETO, sobresalen en la clasificación de mensajes de odio, logrando las mejores métricas de rendimiento y obteniendo el segundo lugar en la competencia HOMO-MEX. Sin embargo, el modelo mDeBERTaV3, a pesar de su arquitectura avanzada y multilingue, mostró un menor desempeño en algunos casos, acercándose al modelo de menor rendimiento, Multinomial Naive Bayes. En contraste, Linear SVC, a pesar de ser un modelo estadísticamente más simple y con menor costo computacional, logró un rendimiento notable, especialmente al ajustarse con matrices TF-IDF. Esto demostró que para textos cortos que contienen términos LGTBI-fóbicos, características como el orden de las palabras y la similitud semántica no son tan relevantes. Este hallazgo sugiere que Linear SVC podría ser una alternativa eficiente y económica frente a los Transformers para clasificar contenido LGTBI-fóbico en redes sociales.

En la tesis de maestría titulada *"Hate Speech Detection Using Neural Networks"* (Spithas, 2020), se propone el desarrollo de una herramienta automatizada diseñada para detectar discurso de odio en redes sociales, con énfasis en Twitter. El objetivo es identificar y clasificar contenido que promueva odio hacia grupos o individuos por aspectos como raza, religión y orientación sexual. Para este propósito, los autores utilizan redes neuronales, específicamente redes LSTM, para el procesamiento del lenguaje natural. Las evaluaciones indican que los modelos que utilizan embeddings GloVe superan a aquellos que generan sus propios embeddings, lo cual sugiere que las representaciones preentrenadas son más efectivas para capturar los matices semánticos relevantes para esta tarea. Sin embargo, algunos modelos clasificaron incorrectamente discursos neutrales como discurso de odio, problema atribuido a un desbalance en el conjunto de datos, donde el discurso de odio representaba aproximadamente el 80% del total.

Revisiones sistemáticas

Preguntas de investigación

En la fase inicial de este mapeo sistemático, se definieron las siguientes preguntas de investigación (Q) y sus respectivas motivaciones (M):

Pregunta 1: ¿Cuáles son las aplicaciones actuales de las redes neuronales en la identificación de discurso de odio en español, y cómo están impactando el desarrollo de sistemas de detección automática?

Motivación: En los últimos años, la inteligencia artificial ha avanzado notablemente, demostrando su papel fundamental en el control del discurso de odio en plataformas digitales y redes sociales. Sin embargo, es crucial comprender cómo se están implementando los modelos de redes neuronales en la clasificación de textos en español y evaluar su impacto real en la moderación de contenido.

Pregunta 2: ¿Cómo se han utilizado las redes neuronales para mejorar la detección del discurso de odio, y cuál es su impacto en el análisis automático de textos?

Motivación: La implementación de redes neuronales para la detección de contenido de odio en plataformas digitales ha traído tanto ventajas como desafíos significativos en el corto y largo plazo. Identificar y comprender su efectividad, limitaciones y potencial nos permitirá evaluar críticamente el impacto real de estas tecnologías en el análisis automático de textos.

Pregunta 3: ¿Cómo puede aprovecharse la inteligencia artificial para optimizar el análisis estilométrico de textos en español, y cuáles son las implicaciones para la detección de discurso de odio?

Motivación: Explorar cómo se utiliza la inteligencia artificial para adaptar modelos de redes neuronales en la detección de discurso de odio en español mediante el uso de características estilométricas nos permitirá comprender en detalle las estrategias de moderación de contenido digital y sus implicaciones en la identificación de discursos ofensivos.

Métodos de revisión

Para realizar este estudio, se utilizó la metodología de Mapeo Sistemático de la Literatura, que permitió realizar una búsqueda exhaustiva de información en artículos científicos sobre temas específicos. Primero, se identificó un problema y se estableció como foco central de la investigación. Luego, se abordaron rigurosamente estos problemas con base en el estado actual de la literatura. Este enfoque permitió determinar el alcance del campo de investigación y combinar diversos aspectos para responder las preguntas planteadas. Al concluir las búsquedas, se derivaron conclusiones a partir de los resultados obtenidos.

Fuentes

Para reunir los artículos necesarios para llevar a cabo el mapeo sistemático de la literatura, se utilizaron amplias fuentes electrónicas, seleccionando bases de datos con documentos de diversas áreas de investigación. Por ello, se eligieron varias de las fuentes más reconocidas con la mayor cantidad de artículos relacionados con la IA (inteligencia artificial), específicamente el uso de redes neuronales para la detección de odio en textos y todos los aspectos relacionados con la temática central del estudio. Las fuentes consultadas incluyeron IEEE Xplore, Science Direct, arXiv, RUC y UTAR, siendo Google Scholar la fuente principal utilizada.

Definición de términos

Se propusieron dos cadenas de búsqueda (SS). La primera cadena utilizó cuatro términos principales: inteligencia artificial, redes neuronales, discurso de odio y textos en español. Esta primera cadena generó SS1: (*"Inteligencia Artificial" AND "Redes Neuronales" AND "Discurso de Odio" AND "Textos en Español"*). La aplicación de esta cadena de búsqueda dio como resultado 12 artículos, lo que

representó una muestra limitada para realizar una revisión sistemática adecuada. Por ello, se propuso una segunda cadena de búsqueda usando términos en inglés para obtener más artículos relacionados con el tema. La segunda cadena generó SS2: (*"Artificial Intelligence" AND "Neural Networks" AND "Hate Speech" AND "Spanish Texts"*). La aplicación de esta segunda cadena resultó en 51 artículos, lo que proporcionó una muestra suficiente para el análisis propuesto.

Estudios incluidos y excluidos

Tras realizar las búsquedas, se aplicaron criterios de inclusión (CI) y de exclusión (CE) a todos los artículos encontrados. Se seleccionaron aquellos que contenían información relevante para responder a las preguntas del estudio y se descartaron los que no cumplían con los requisitos. Este proceso fue iterativo, aplicando diferentes reglas de selección a cada artículo. Es importante señalar que, en esta etapa, se obtuvo un conjunto inicial de artículos relevantes que posteriormente fueron sometidos a un análisis detallado mediante la aplicación de filtros usando los criterios establecidos. La selección final de los trabajos se completó utilizando los criterios de inclusión y exclusión, los cuales se detallan en la Tabla 1.

Tabla 1. Criterios de selección y exclusión.

Código	Descripción
Criterios de inclusión	
CI1	Artículos publicados entre 2018 y 2024.
CI2	Artículos que relacionan inteligencia artificial, redes neuronales, discurso de odio y textos en español.
CI3	Tipos de documentos: tesis, artículos científicos, ponencias, revistas y libros.
Criterios de exclusión	
CE1	Artículos publicados antes de 2018.
CE2	Artículos que no consideran el análisis de discurso de odio mediante redes neuronales como tema primario o secundario.
CE3	Artículos que no mencionan el uso de texto en español.
CE4	Artículos duplicados.

Fuente: Elaboración propia

Utilizando la segunda cadena de búsqueda (SS2), se encontraron 51 estudios. Tras aplicar los criterios de inclusión y exclusión, se seleccionaron 50 estudios, de los cuales 15 fueron identificados como candidatos potenciales para este análisis.

Resultados y discusión

A través de la búsqueda y selección de artículos se reunió un conjunto representativo que muestra cómo ha evolucionado el tema de investigación durante el período establecido en el protocolo. Esto facilitó la extracción de información y la identificación de técnicas y tecnologías utilizadas para abordar las preguntas de investigación. A continuación, se presenta un resumen de los resultados obtenidos en el mapeo sistemático de la literatura.

Pregunta 1: Cuáles son las aplicaciones actuales de las redes neuronales en la identificación de discurso de odio en español, y cómo están impactando el desarrollo de sistemas de detección automática?

En el artículo *"A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis"* se realiza un estudio que destaca las aplicaciones específicas de las redes neuronales en la identificación de discurso de odio en español, incluyendo la integración del análisis de polaridad y emociones para mejorar la precisión en la detección de contenido dañino. Por ejemplo, el enfoque multitarea propuesto utiliza el modelo BETO, una versión en español de BERT, para transferir conocimiento entre tareas relacionadas, mejorando la identificación de patrones afectivos compartidos y, en consecuencia, la detección de discurso de odio en redes sociales. Adicionalmente, la metodología incorpora el uso de conjuntos de datos específicos de Twitter y técnicas de preprocesamiento adaptadas a esta plataforma, lo que fortalece la aplicabilidad y efectividad del modelo en contextos del mundo real (Plaza-del Arco et al., 2021).

En el artículo *"Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data"*, el estudio examina cómo las redes neuronales pueden aplicarse a la detección automática de expresiones sexistas en redes sociales, específicamente en Twitter. Esta investigación emplea modelos de aprendizaje profundo, como Redes Neuronales Convolucionales (CNN) y Redes de Memoria a Corto y Largo Plazo (LSTM), junto con técnicas de aprendizaje por transferencia y representaciones de palabras preentrenadas, con el objetivo de detectar tanto el sexismo explícito como el sutil en los textos. Las aplicaciones actuales de las redes neuronales para la detección de discurso de odio en español se centran en identificar palabras ofensivas y analizar el sentimiento en redes sociales, ayudando a capturar no solo el odio manifiesto, sino también comportamientos dañinos más sutiles que afectan la libertad de expresión y la convivencia en línea. Estos avances están influyendo significativamente en el desarrollo de sistemas de detección automática, ya que permiten que las plataformas digitales comprendan mejor los matices del lenguaje ofensivo y tomen las medidas adecuadas (Rodríguez-Sánchez et al., 2020).

Pregunta 2: Cómo se han utilizado las redes neuronales para mejorar la detección de discurso de odio, y cuál es su impacto en el análisis automático de textos?

En el artículo *"Detecting Offensive Language by Integrating Multiple Linguistic Phenomena"* se propone un enfoque novedoso que utiliza redes neuronales para abordar la difícil tarea de identificar el discurso de odio en español. Mediante el uso de modelos avanzados de aprendizaje profundo, como BETO, una variante de BERT específicamente adaptada para el español, se exploran técnicas como Transfer Learning y Multi-Task Learning (MTL), lo que permite una comprensión más profunda de fenómenos linguísticos como el sarcasmo, las emociones y las figuras retóricas. Esta integración de fenómenos linguísticos mejora significativamente la precisión de los sistemas en la identificación de discurso de odio, sexismo y misoginia en plataformas digitales. El estudio demuestra que el uso de estos modelos de redes neuronales no solo permite una detección más precisa de contenido dañino,

sino que también contribuye significativamente al análisis automático de textos, permitiendo que las herramientas de procesamiento del lenguaje natural analicen mejor el contexto y la intención detrás de los comentarios (Plaza-del Arco, 2023).

En el artículo *"Detecting Racism and Xenophobia Using Deep Learning Models on Twitter Data: CNN, LSTM, and BERT"* se propone un modelo para la detección automatizada de mensajes racistas y xenófobos en español, un área crucial dada la escasez de modelos entrenados en este idioma. Para mejorar la precisión en la clasificación de este tipo de contenido se implementaron redes neuronales avanzadas como CNN, LSTM y, en particular, el modelo BETO, una variante de BERT adaptada exclusivamente para el español. BETO mostró una precisión superior al 85%, superando a otros modelos, incluidos los multilingues como mBERT, gracias a su capacidad para capturar los matices y particularidades linguísticas del español. La aplicación de estas redes neuronales permite una clasificación y análisis de contenido más efectivos, automatizando un proceso que sería inviable manualmente dada la enorme cantidad de datos generados diariamente en redes sociales (Benítez-Andrades et al., 2022).

Pregunta 3: Cómo puede aprovecharse la inteligencia artificial para optimizar el análisis estilométrico de textos en español, y cuáles son las implicaciones para la detección de discurso de odio?

En el artículo *"Classifiers and Text Mining: Application to a Specific Context"* se realiza un estudio que explora el uso de la inteligencia artificial (IA) para mejorar el análisis estilométrico de textos en español, centrándose en la detección de discurso sexista en redes sociales. Este trabajo examina cómo modelos avanzados de procesamiento del lenguaje natural (NLP), como BERT y su variante en español, SpanBERTa, pueden reconocer patrones linguísticos y contextuales en los textos, facilitando la clasificación de contenido dañino. Aprovechando la capacidad de la IA para procesar grandes volúmenes de datos y detectar matices del lenguaje, este estudio demuestra la viabilidad de identificar con precisión el lenguaje ofensivo y sexista en múltiples idiomas. Las implicaciones de esta tecnología son significativas: su aplicación efectiva en español abre el camino para una moderación más eficiente en redes sociales, permitiendo la detección y clasificación automática de comportamientos y discursos de odio con alta precisión (Santos Ríos, 2022).

En el artículo *"Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review"*, el estudio explora cómo los modelos de word embedding y las técnicas de aprendizaje profundo pueden utilizarse para optimizar el análisis estilométrico de textos en español, centrándose en la detección de discurso de odio. Al analizar modelos como Word2Vec, GloVe, BERT y GPT, la investigación demuestra cómo las representaciones contextuales y semánticas mejoran la precisión en tareas de clasificación de sentimientos y detección de contenido ofensivo. Estos modelos capturan las complejidades linguísticas y culturales del idioma, lo que es crucial en el español, donde el contexto juega un papel clave en la identificación de contenido dañino. Al aplicar técnicas avanzadas de procesamiento del lenguaje natural (NLP), los hallazgos de este estudio sugieren que no solo es posible identificar el discurso de odio con mayor precisión, sino también adaptar estos modelos a entornos multilingues, ampliando su utilidad para la moderación de contenido en redes sociales (Asudani et al., 2023).

Metodología

El diseño metodológico de la investigación propuesta se basa en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta es una metodología popular para incrementar el éxito de los proyectos de minería de datos. Establece una serie flexible de seis etapas que facilitan la creación y aplicación de un modelo de minería de datos para su uso en un entorno práctico. Esta metodología define un proyecto como un proceso cíclico, donde pueden utilizarse múltiples iteraciones para garantizar que el resultado final se alinee mejor con los objetivos del negocio. Después de identificar la meta a alcanzar (fase de Comprensión del Negocio), los datos deben analizarse (Comprensión de los Datos) y procesarse (Preparación de los Datos). Los datos contienen conceptos (lo que debe aprenderse), instancias (registros independientes relacionados con una ocurrencia) y atributos (que caracterizan un aspecto específico de una instancia dada). La fase de Modelado construye el modelo que representa el conocimiento aprendido. Si el modelo resultante no es lo suficientemente bueno para soportar el negocio, se define una nueva iteración de CRISP-DM. De lo contrario, el modelo se implementa (Moro et al., 2011).

La metodología consta de seis fases:

- Comprensión del Negocio
- Comprensión de los Datos
- Preparación de los Datos
- Modelado
- Evaluación
- Despliegue

Es importante señalar que la fase de despliegue no se implementará en este trabajo, ya que no se ejecutará en un entorno de producción.

Comprensión del negocio

En esta primera fase, se ha establecido una base sólida para la creación de un modelo de detección de discurso de odio basado en redes neuronales. Para este propósito, se utilizaron dos conjuntos de datos: uno de la plataforma Kaggle, denominado "Spain_test.csv" (Moosa and Najiba, 2022), y otro de la plataforma Hugging Face, denominado "es_hf_102024.csv" (Tonneau et al., 2024). Tras su análisis, fueron unificados siguiendo una estructura estandarizada, generando dos nuevos conjuntos de datos: "speeches_no_hate.csv" y "speeches_hate.csv". El objetivo es entrenar y evaluar cuatro modelos de redes neuronales (CNN, RNN, LSTM y Transformer) para la detección de discurso de odio en textos en español, analizar sus resultados, comprender el comportamiento de los modelos, generar un modelo de mayor desempeño y evaluar la usabilidad y efectividad de estas redes para la moderación de contenido ofensivo en plataformas digitales.

Comprensión de los datos

La metodología CRISP-DM considera la "Exploración y Comprensión de los Datos" como su segunda etapa. Durante esta fase se realizó la recolección inicial de los datos que se utilizarán para entrenar y evaluar los modelos de redes neuronales en la detección de discurso de odio en textos en español. Esta fase es crucial para establecer un primer contacto con el problema, familiarizarse con la naturaleza de los datos, evaluar su calidad e identificar las relaciones más evidentes entre las variables.

Los datos de entrenamiento del conjunto *speeches_no_hate.csv* están conformados por las siguientes columnas:

- **Tweet_id:** Contiene los IDs de los usuarios que generaron comentarios en la plataforma.
- **Speeches:** Contiene los comentarios de los usuarios en plataformas digitales etiquetados como "no odio", utilizados para el entrenamiento del modelo.
- **Label:** Indicador para identificar los comentarios etiquetados como "no odio" (0).

Los datos de entrenamiento del conjunto *speeches_hate.csv* están conformados por las siguientes columnas:

- **Tweet_id:** Contiene los IDs de los usuarios que generaron comentarios en la plataforma.
- **Negative_speeches:** Contiene los comentarios etiquetados como "odio", utilizados para el entrenamiento del modelo.
- **Label:** Indicador para identificar los comentarios etiquetados como "odio" (1).

Preparación de los datos

La metodología CRISP-DM considera la "Preparación de los Datos" como su tercera etapa. En esta fase, la información se procesa y limpia para asegurar que sea adecuada para su uso en las técnicas de minería de datos aplicadas en el estudio sobre la identificación de discurso de odio en textos en español. Este proceso es esencial para garantizar la calidad y confiabilidad de los datos, los cuales serán utilizados posteriormente para entrenar y evaluar los modelos de redes neuronales.

Las características estilométricas seleccionadas para esta investigación fueron las siguientes:

- **Longitud promedio de palabra:** Calcula la longitud promedio de las palabras en un texto. Para obtener una medida precisa, el texto se preprocesa primero eliminando los signos de puntuación y convirtiéndolo a minúsculas. El cálculo se realiza dividiendo el número total de caracteres entre el número total de palabras, proporcionando un promedio de cuántos caracteres contiene cada palabra del texto.
- **Diversidad léxica:** Mide la riqueza y variedad del vocabulario empleado en un texto. Generalmente se calcula mediante la razón entre el número de palabras únicas y el número total de palabras, indicando cuán variado es el lenguaje empleado.
- **Longitud promedio de oración:** Mide el número promedio de palabras por oración en un texto. Para realizar este cálculo, se utiliza primero la función `sent_tokenize` para identificar y separar automáticamente las oraciones del texto. Luego, para cada oración, se cuenta el número de palabras y, finalmente, se calcula el promedio de estas longitudes utilizando la función `mean` de NumPy. Esta medida ayuda a evaluar la complejidad sintáctica y el estilo de escritura del texto.
- **Número de oraciones:** Cuantifica el número total de oraciones en un texto. Sirve como indicador básico de la estructura y longitud del documento, brindando información sobre la fragmentación y complejidad sintáctica.
- **Longitud total del documento:** Mide la longitud completa del texto, contando el número total de caracteres en todas las oraciones. Esta métrica proporciona una visión cuantitativa del tamaño y la dimensión global del documento.
- **Puntuación de contenido ofensivo:** Calcula el porcentaje de palabras ofensivas presentes en un texto en relación con el número total de palabras. El proceso comienza leyendo un conjunto predefinido de palabras ofensivas desde un archivo CSV almacenado en una carpeta de Drive, el cual contiene términos y frases adaptadas al contexto ecuatoriano. Estos términos se obtuvieron de una tesis que abordó temas relacionados con la comunicación (Esteve Bañón, 2020). Para realizar el cálculo, el texto se normaliza convirtiéndolo a minúsculas y separándolo en palabras individuales. Luego, se identifica cuántas de estas palabras coinciden con las del conjunto predefinido. Los resultados se expresan como porcentajes, donde un valor más alto indica una mayor proporción de palabras ofensivas en el texto.

Para la preparación de los datos de los comentarios de los usuarios, primero se importa la librería pandas para el manejo y análisis de datos en Python. Se cargan dos conjuntos de datos de discursos de usuarios (`discursos_no_odio.csv` y `discursos_odio.csv`). Solo se seleccionan las columnas *speeches* y *speeches_negative*, correspondientes a las variables `dataset_no_hate` y `dataset_hate`. La función `dropna()` se utiliza para eliminar las filas con valores nulos o vacíos. Adicionalmente, se asigna la etiqueta '1' a los comentarios con contenido de odio (`dataset_hate`) y la etiqueta '0' a los comentarios neutrales (`dataset_no_hate`).

El proceso comienza con la concatenación de los dos conjuntos de datos: el dataset que contiene textos clasificados como discurso de odio (`dataset_hate`) y aquellos clasificados como no odio (`dataset_no_hate`). Una vez consolidado el conjunto, se realiza la extracción de las características estilométricas de los textos. Estas características representan los patrones linguísticos y estilísticos que servirán como base para el entrenamiento de los modelos. Posteriormente, los datos se separan en características y etiquetas. Por un lado, la matriz de características `X` se construye eliminando las columnas `tweet_id` y `label` del DataFrame, conservando únicamente

los atributos estilométricos. Por otro lado, se extrae el vector de etiquetas y, que almacena la clasificación binaria de cada texto (donde 0 representa la ausencia de discurso de odio y 1 indica su presencia).

Los datos concatenados se dividen en dos conjuntos para entrenamiento (`X_train` y `y_train`) y prueba (`X_test` y `y_test`):

- **X_train:** Conjunto de entrenamiento que contiene el 80% de los datos de X (características estilométricas).
- **X_test:** Conjunto de prueba que contiene el 20% de los datos de X (características estilométricas).
- **y_train:** Conjunto de entrenamiento que contiene el 80% de los datos de y (etiquetas) correspondientes a X_train.
- **y_test:** Conjunto de prueba que contiene el 20% de los datos de y (etiquetas) correspondientes a X_test.

La división se realiza mediante `train_test_split` con un `test_size=0.2` y `random_state=42`, tal como se muestra en la Ecuación (1):

$$(X_{train}, X_{test}, y_{train}, y_{test}) = \text{split}(X, y, 0.2, 42) \quad (1)$$

Modelado

El desarrollo del modelado es una parte clave al trabajar con datos siguiendo el método CRISP-DM. En esta etapa se seleccionan y aplican las técnicas de modelado más apropiadas para los datos y el problema que se busca resolver. En este caso, implica la creación de modelos de redes neuronales para detectar discurso de odio en textos en español, utilizando características estilométricas.

Para el procesamiento de las expresiones de los usuarios, se implementaron cuatro modelos de redes neuronales (CNN, RNN, LSTM y Transformer). En la configuración de cada modelo se establecieron hiperparámetros específicos, incluyendo el tamaño del lote para el procesamiento por lotes, la tasa de aprendizaje que controla la magnitud de las actualizaciones de pesos y la tasa de dropout para prevenir el sobreajuste durante el entrenamiento. Los hiperparámetros clave son:

- **Tamaño de lote (Batch size):** Se han definido 128 muestras para un procesamiento eficiente de los datos. Este valor se considera una práctica estándar en aprendizaje profundo, ya que proporciona un equilibrio entre la estabilidad del entrenamiento y la capacidad de generalización del modelo.
- **Tasa de aprendizaje (Learning rate):** Se ha fijado en 0,001 para el optimizador Adam, ya que este valor proporciona un balance adecuado entre velocidad y precisión de aprendizaje, permitiendo que el modelo ajuste sus pesos de forma gradual y estable hacia el óptimo global sin oscilaciones excesivas.
- **Tasa de dropout:** Se ha establecido entre 0,2 (20%) y 0,5 (50%) para las capas densas del modelo, como medida esencial para prevenir el sobreajuste durante el entrenamiento.

Modelado de la red neuronal – CNN (*Convolutional Neural Network*)*:* En este modelado se implementan convoluciones en paralelo con diferentes tamaños de filtro (2, 3 y 4) para capturar patrones a distintas escalas en la entrada. Luego, cada salida convolucional se pasa por una capa GlobalMaxPooling1D que extrae la información más relevante, y estas salidas se concatenan para formar una representación combinada. Finalmente, se añaden capas densas con activación ReLU y regularización L2, junto con Dropout, para refinar las características y mitigar el sobreajuste.

Modelado de la red neuronal – RNN (*Recurrent Neural Network*)*:* Se implementa una red neuronal recurrente compuesta por dos capas SimpleRNN en secuencia. La primera capa, con 128 neuronas y función de activación tanh, devuelve secuencias (`return_sequences=True`) para permitir el procesamiento continuo de información temporal y emplea regularización L2 para controlar el sobreajuste. Posteriormente se aplica Dropout para reducir la correlación entre neuronas y mejorar la generalización. Luego se añade otra capa SimpleRNN con 64 neuronas, nuevamente con regularización L2 y Dropout. Finalmente se incorpora una capa densa de salida con activación sigmoide.

Modelado de la red neuronal – LSTM (*Long Short-Term Memory*)*:* En este modelado se utilizan capas LSTM bidireccionales en tres niveles (con 96, 64 y 32 unidades, respectivamente) para capturar características temporales a diferentes profundidades, cada una acompañada de Dropout y regularización L2. Simultáneamente, se recibe una segunda entrada con características de estilo, que se procesa mediante una capa densa y Dropout. Ambas salidas se combinan usando Concatenate, y la fusión resultante pasa por otra capa densa regularizada antes de generar la salida final, consistente en una neurona con activación sigmoide.

Modelado de la red neuronal – Transformer: En este modelado se utiliza un Transformer para extraer características relevantes de una entrada principal, a la cual se le aplica primero un embedding denso y un reshape. Luego se repiten varios bloques Transformer (TransformerBlock), cada uno con múltiples cabezas de atención (multi-head attention), una capa feed-forward de dimensión ajustable y dropout para regularización. Tras aplanar la salida del Transformer, se integra una segunda rama para procesar características adicionales, consistente en una capa densa con activación ReLU y dropout. Finalmente, ambas salidas se concatenan y pasan por capas densas decrecientes, aplicando dropout para mitigar el sobreajuste, hasta llegar a la capa de activación sigmoide final.

Resultados

Tras completar la etapa de creación de los modelos, esta fase tiene como objetivo evaluar y seleccionar la red neuronal con mejor desempeño en la detección de discurso de odio en textos en español, utilizando características estilométricas. Para ello se entrenaron las cuatro redes neuronales utilizando métricas estándar de clasificación: precisión, recall y F1-score.

En la preparación de datos se incluyen etiquetas donde la Clase 0 representa los discursos neutrales o positivos, mientras que la Clase 1 representa los discursos de odio.

Con esta aclaración, se realiza una evaluación exhaustiva de cada modelo en la Tabla 2 para identificar la opción que mejor cumpla con los requisitos establecidos.

Evaluación del modelo CNN: La precisión para la clase 0 es moderadamente baja (0,80), mientras que su recall es alto (0,90), lo que indica que el modelo tiene una capacidad moderada para detectar correctamente los discursos neutrales o positivos. Por otro lado, la clase 1 muestra mejor desempeño, con alta precisión y recall (0,92 y 0,85, respectivamente), lo que resulta en un F1 elevado (0,88). El modelo logró una precisión global del 87%, con una macro-precisión del 86% y un recall del 87%, demostrando un equilibrio consistente entre precisión y recall para ambas clases.

Evaluación del modelo RNN: La precisión para la clase 0 es moderadamente baja (0,79), mientras que su recall es alto (0,91), lo que indica una capacidad moderada para detectar correctamente los discursos neutrales o positivos. La clase 1 muestra mejor desempeño, con alta precisión y recall (0,93 y 0,84, respectivamente), lo que da como resultado un F1 alto (0,88). El modelo alcanzó una precisión global del 87%, con una macro-precisión del 86% y un recall del 87%.

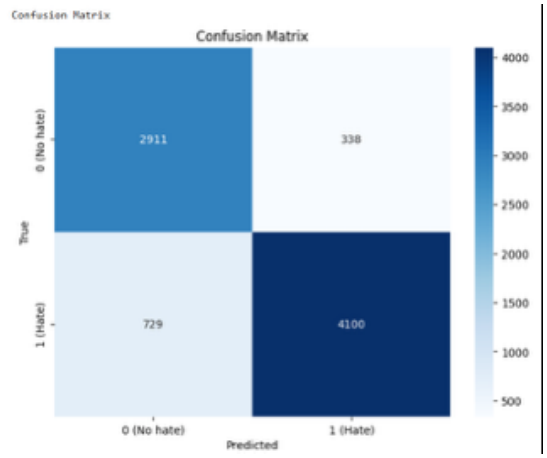
Evaluación del modelo LSTM: La precisión y el recall muestran valores relativamente altos (por encima del 80%), lo que indica que el modelo tiene buena capacidad para detectar correctamente los discursos neutrales o positivos. La clase 1 muestra un desempeño superior, con precisión y recall por encima del 85%, alcanzando un F1 de 0,88. El modelo logró una precisión global del 86%, con una macro-precisión del 85% y un recall del 86%.

Evaluación del modelo Transformer: La precisión para la clase 0 es moderadamente baja (0,80), mientras que su recall es alto (0,93). La clase 1 muestra un mejor desempeño, con muy buena precisión y recall (0,95 y 0,84, respectivamente), lo que resulta en un F1 alto (0,89). El modelo logró una precisión global del 88%, con una macro-precisión del 87% y un recall del 88%.

Las matrices de confusión son una herramienta fundamental en aprendizaje automático para evaluar la precisión de los modelos de clasificación cuando existen dos posibles verdades: el discurso es o no es de odio. La matriz de confusión organiza estos resultados en una tabla que facilita la evaluación del desempeño de los modelos de clasificación (CNN, RNN, LSTM y Transformer). Donde los verdaderos positivos y los verdaderos negativos indican predicciones correctas, y los falsos positivos y falsos negativos indican errores del modelo. Esta herramienta es crucial para comprender la precisión general del modelo y los tipos específicos de errores que comete. A continuación, se muestran las matrices de confusión de cada modelo en las Figuras 1–4.

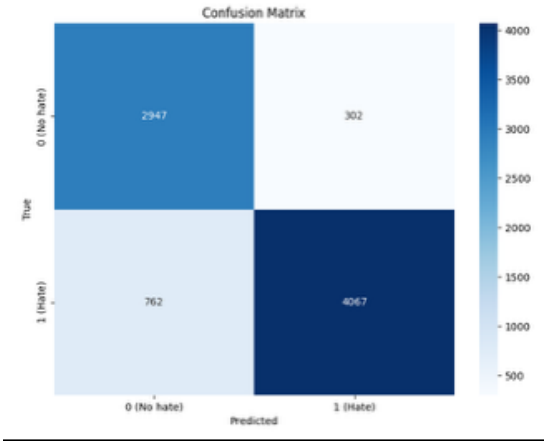
Conclusiones

Este estudio ha demostrado la efectividad de las redes neuronales en el análisis automatizado de discurso para la detección de odio en textos en español, con el objetivo de moderar contenido ofensivo en entornos digitales en Ecuador. La comparación entre los modelos CNN, RNN, LSTM y Transformer reveló diferencias significativas en su desempeño:



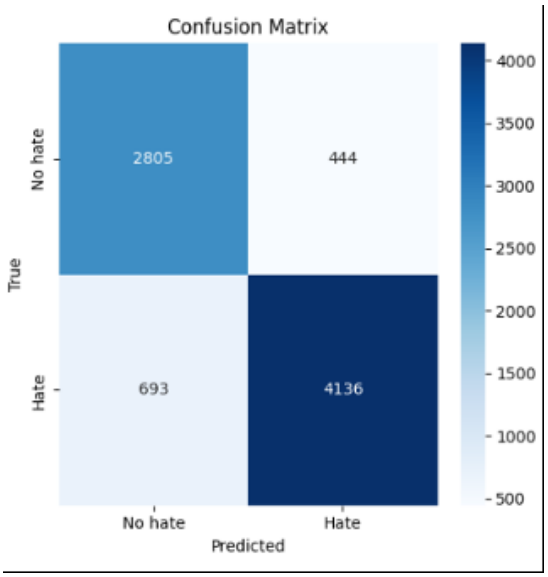
Fuente: Elaboración propia

Figura 1. Matriz de confusión del modelo CNN.



Fuente: Elaboración propia

Figura 2. Matriz de confusión del modelo RNN.



Fuente: Elaboración propia

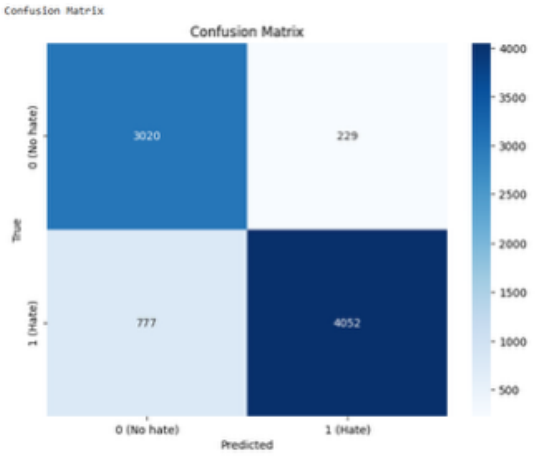
Figura 3. Matriz de confusión del modelo LSTM.

- El modelo CNN presenta un balance notable entre precisión y recall para ambas clases, reflejado en un AUC de 0,92. Sin embargo, a pesar de su robustez,

Tabla 2. Evaluación de los modelos.

Modelo	Clase	Precisión	Recall	F1-Score	Soporte
CNN	0 (No odio)	0,80	0,90	0,85	3249
	1 (Odio)	0,92	0,85	0,88	4829
	Total	**0,87**	—	—	**8078**
RNN	0 (No odio)	0,79	0,91	0,85	3249
	1 (Odio)	0,93	0,84	0,88	4829
	Total	**0,87**	—	—	**8078**
LSTM	0 (No odio)	0,80	0,86	0,83	3249
	1 (Odio)	0,89	0,86	0,88	4829
	Total	**0,85**	—	—	**8078**
Transformer	0 (No odio)	0,80	0,93	0,86	3249
	1 (Odio)	0,95	0,84	0,89	4829
	Total	**0,87**	—	—	**8078**

Fuente: Elaboración propia



Fuente: Elaboración propia

Figura 4. Matriz de confusión del modelo Transformer.

no destaca como el mejor en todas las métricas, por lo que no alcanza la calificación más alta entre los modelos evaluados.

- El modelo RNN logra la misma precisión global que la CNN (87%) y demuestra un desempeño general estable. No obstante, su AUC de 0,91 es ligeramente inferior a los mejores resultados, lo que sugiere un desbalance moderado en la clasificación de clases.
- El modelo LSTM alcanza la menor precisión global (86%) entre las cuatro arquitecturas. Aunque su F1 ponderado (0,86) sigue siendo aceptable, queda rezagado respecto a los demás modelos, por lo que no se considera la mejor alternativa en términos de desempeño general.
- El modelo Transformer se destaca como la alternativa más eficiente. Si bien su precisión (87%) coincide con la de CNN y RNN, posee un F1 ponderado superior (0,88) y un AUC de 0,92, lo que indica un balance óptimo entre ambas clases. Adicionalmente, el bajo número de falsos positivos confirma su efectividad para identificar la clase 0 (no odio), mientras que la alta tasa de verdaderos positivos demuestra su solidez al clasificar la clase 1 (odio).

Estos resultados demuestran que la arquitectura Transformer, combinada con el uso de características estilométricas, es particularmente eficaz para el análisis de discurso orientado a la detección de mensajes de odio en textos en español, con un enfoque específico en la diversidad linguística de Ecuador. Su implementación podría convertirse en una herramienta esencial en esta era digital, donde el respeto por la libertad de expresión y la promoción de la convivencia pacífica son pilares fundamentales para los ciudadanos ecuatorianos.

Referencias Bibliográficas

Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S. K., and Asghar, M. Z. (2025). Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, 15(1).

Álvarez Benjumea, A. (2025). Social norms and the expression of prejudice: How the norm changes. *Current Opinion in Psychology*, 62.

Assimakopoulos, S. and Serafis, D. (2025). Why soft hate speech matters: argumentativity and the dispersion of hatred towards minorities. *Critical Discourse Studies*.

Asudani, D. S., Nagwani, N. K., and Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, 56(9):10345–10425.

Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J.-M., and García-Ordás, M. T. (2022). Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. *PeerJ Computer Science*, 8:e906.

Esteve Bañón, Z. (2020). *Análisis de las características ambientales del ciberlugar para la detección y prevención de la comunicación violenta y el discurso de odio*. PhD thesis, Universidad Miguel Hernández de Elche. Tesis doctoral no publicada.

Fernández Rosauro, C. (2023). Detección de discurso de odio contra la comunidad LGTBI+ mexicana mediante

el uso de arquitecturas transformer. Master's thesis, Universitat Oberta de Catalunya. Tesis de maestría no publicada.

Fu, L., Zhang, X., Liao, C., Xiong, P., Song, N., and Chen, K. (2025). Microtia and social media: can patients find reliable information in TikTok? *European Journal of Plastic Surgery*, 48(1).

Gonçalves, M., Salgado, C., de Sousa, A., and Teixeira, L. (2025). Data storytelling and decision-making in seaport operations: A new approach based on business intelligence. *Sustainability*, 17(1).

Hassan, M. M., Orebi, H. A., Salama, B., and Kabbash, I. A. (2025). The association of social media use and other social factors with symptoms of attention-deficit/hyperactivity disorder in Egyptian university students. *BMC Psychiatry*, 25(1).

Jiménez-Díaz, O. and Del Rey, R. (2025). Cyberhate in adolescents and youths: A systematic review of labels and associated factors. *Aggression and Violent Behavior*, 81.

Miró Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en internet. *IDP. Revista de Internet, Derecho y Polítca*, (22).

Moosa, W. H. and Najiba (2022). Multi-lingual HateSpeech dataset. Kaggle. Conjunto de datos.

Moro, S., Laureano, R., and Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *Proceedings of the European Simulation and Modelling Conference* (*ESM*).

Park, J. R., Poole, E., and Li, J. (2024). Stylometric features in librarian's responses to user queries: implications for user interaction in digital information services. *Global Knowledge, Memory and Communication*, 73(3):375–390.

Patel, D., Pramanik, P. K. D., Suryawanshi, C., and Pareek, P. (2025a). Detecting toxic comments on social media: an extensive evaluation of machine learning techniques. *Journal of Computational Social Science*, 8(1).

Patel, D., Pramanik, P. K. D., Suryawanshi, C., and Pareek, P. (2025b). Machine learning techniques for hate speech detection: a comprehensive evaluation. *Journal of Computational Social Science*, 8(1).

Perry, B. and Olsson, P. (2009). Cyberhate: The globalization of hate. *Information and Communications Technology Law*, 18(2).

Plaza-del Arco, F. M. (2023). *Detecting offensive language by integrating multiple linguistic phenomena*. PhD thesis, Universidad de Jaén. Tesis doctoral no publicada.

Plaza-del Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., and Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 8:219563–219576.

Romero-Vega, R. R., Cumbicus-Pineda, O. M., López-Lapo, R. A., and Neyra-Romero, L. A. (2021). Detecting xenophobic hate speech in Spanish tweets against Venezuelan immigrants in Ecuador using natural language processing. In *Communications in Computer and Information Science*, volume 1388. Springer.

Santos Ríos, R. (2022). *Classifiers and text mining: application to a specific context*. PhD thesis, Universidad de Salamanca. Tesis doctoral no publicada.

Shohan, M. H., Hossain, M., Rahman, M., and Karim, A. (2025). Use of natural language processing for the detection of hate speech on social media. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 51(2):86–96.

Spithas, V. S. (2020). Hate speech detection using neural networks. Master's thesis, University of Piraeus. Tesis de maestría no publicada.

Suntaxi, J. G., Recalde, L., Loza-Aguirre, E., and Martinez-Mosquera, D. (2024). Cholo, indio, mono: Ecuador and the linguistic context of hate speech on Twitter. In *International Conference on eDemocracy and eGovernment* (*ICEDEG*).

Suntaxi Recalde, J. G. (2022). Modelo de detección de discurso de odio en Ecuador mediante clasificación supervisada de tweets y técnicas de NLP. Master's thesis, Escuela Politécnica Nacional, Quito, Ecuador. Tesis de maestría no publicada.

Tonneau, M., Liu, D., Fraiberger, S., Schroeder, R., Hale, S. A., and Rottger, P. (2024). From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms* (*WOAH*).

Ureña-López, L. A., Plaza-del Arco, F. M., Molina-González, M. D., and Martín-Valdivia, M. T. (2024). MODERATES: MODERAtion of ConTEnts in social networks using language technologies. In *CEUR Workshop Proceedings*, pages 112–117.

Zapatier, L., Morejón, M., Escobar, S., Jima-González, A., and Cuenca, E. (2024). Hate speech detection in the 2022 Ecuador strike using the k-nearest neighbors algorithm. In *Lecture Notes in Networks and Systems*, volume 775. Springer.