Hate Speech Identification in Texts Through Phraseological Analysis and TF-IDF Representation of N-Grams

César Espin-Riofrio, MSc.[1], Ángela Yanza-Montalván, Ph.D.[1], Rocío Carchi-Encalada, MGs.[1], Mayra Magdalena Arias Candelario, MGs.[1], Angélica Cruz-Chóez, Ph.D.[1], Juan Montesdeoca-Rodríguez, Ing.[1], Marcos Bailón-Guaranda, Ing.[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, angela.yanzam@ug.edu.ec, rocio.carchie@ug.edu.ec, mayra.magdalenaa@ug.edu.ec, angelica.cruzc@ug.edu.ec, juan.montesdeocar@ug.edu.ec, marcos.bailong@ug.edu.ec

Abstract*– The phenomenon of hate speech, which is widely present on digital platforms, poses unique challenges in the Spanish language due to its linguistic richness and cultural diversity—features that complicate the automatic identification of such content. This complexity is further intensified by the ability of language to disguise hateful messages through sarcasm, irony, or culturally specific references. The present study focuses on the extraction of phraseological features and TF-IDF n-grams, employing traditional classification models based on statistical methods and neural networks, as well as ensemble techniques to enhance overall classification performance. The OffendEs dataset, specifically labeled for hate speech detection tasks in Spanish, was used for training and evaluation. The results show that ensemble models achieve higher accuracy levels, demonstrating a balanced performance across classes and a notable capacity to handle the linguistic complexity of Spanish. In particular, the Voting Classifier achieved a macro F1 score of 0.742261. The results were compared with predictions generated by specialized models trained for hate speech detection, such as Piuba and Pysentimiento, revealing a superior performance of the proposed model, with a 17.84% improvement over Piuba and 8.25% over Pysentimiento. These findings highlight the effectiveness of the proposed methodology and its contribution to the development of more accurate tools for the automatic detection of hate speech in the Spanish language.

***Keywords*-- Hate speech, Phraseological features, n-grams, TF-IDF, Natural Language Processing.**

Identificación de Discurso de Odio en Textos Mediante Análisis Fraseológico y Representación TF-IDF de N-Gramas

César Espin-Riofrio, MSc.[1], Ángela Yanza-Montalván, Ph.D.[1], Rocío Carchi-Encalada, MGs.[1], Mayra Magdalena Arias Candelario, MGs.[1], Angélica Cruz-Chóez, Ph.D.[1], Juan Montesdeoca-Rodríguez, Ing.[1], Marcos Bailón-Guaranda, Ing.[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, angela.yanzam@ug.edu.ec, rocio.carchie@ug.edu.ec, mayra.magdalenaa@ug.edu.ec, angelica.cruzc@ug.edu.ec, juan.montesdeocar@ug.edu.ec, marcos.bailong@ug.edu.ec

Resumen– ***El fenómeno del discurso de odio, ampliamente presente en plataformas digitales, plantea desafíos únicos en el idioma español debido a su riqueza lingüística y diversidad cultural, características que dificultan la identificación automática de este tipo de contenido, lo cual se agrava por la capacidad del lenguaje para disfrazar mensajes de odio mediante el sarcasmo, la ironía o referencias culturales específicas. La presente investigación se enfoca en la extracción de características fraseológicas y n-gramas de TF-IDF, utilizando modelos de clasificación tradicionales basados en estadísticas y redes neuronales, y métodos de ensamblados para repotenciar en conjunto los modelos de clasificación. Se utilizó el dataset OffendEs, etiquetado específicamente para tareas de discurso de odio en español. Los resultados muestran que los modelos ensamblados alcanzan niveles de precisión superiores, logrando un buen equilibrio entre clases y destacando su capacidad para manejar la complejidad lingüística del español, en específico Voting Classifier logró un macro F1 de 0.742261. Los resultados obtenidos fueron comparados con las predicciones generadas por modelos específicos entrenados para la detección de discurso de odio, como Piuba y Pysentimiento, evidenciando un rendimiento superior del modelo propuesto, con una mejora del 17.84 % respecto a Piuba y del 8.25 % en comparación con Pysentimiento. Estos resultados subrayan la efectividad de nuestra metodología y su contribución al desarrollo de herramientas más precisas para la detección automática de discurso de odio en idioma español.***

Palabras clave-- Discurso de odio, características fraseológicas, n-gramas, TF-IDF, Procesamiento de Lenguaje Natural.

I. Introducción

Durante los últimos años, el discurso de odio en textos se ha convertido en un fenómeno cada vez más común, afectando plataformas que van desde los foros públicos hasta los medios de comunicación. Este lenguaje, caracterizado por incitar al odio, la discriminación y la violencia contra grupos específicos, ha despertado un interés creciente en la comunidad científica, particularmente en el ámbito del Procesamiento del Lenguaje Natural (PLN). Aunque se han desarrollado modelos para la detección automática de discurso de odio en varias lenguas, el español plantea desafíos únicos debido a su riqueza lingüística y diversidad cultural, lo que subraya la importancia de investigaciones orientadas a abordar estas complejidades.

Según [1], el discurso de odio se define como todo mensaje público de rechazo o menosprecio dirigido contra grupos sociales caracterizados por su situación actual o potencial de marginación social, o por haber sido tradicionalmente objeto de discriminación. De manera complementaria, [2] lo describe como cualquier comentario destinado a menospreciar, humillar o incitar al odio hacia un grupo o clase de personas.

La llegada de las redes sociales y los foros en línea ha transformado profundamente las dinámicas de comunicación y la expresión de opiniones en la era digital [3], la difusión de discursos de odio en textos escritos puede provocar diversos problemas sociales y psicológicos significativos, afectando tanto a individuos como a comunidades, si no se identifican ni se abordan oportunamente, estas expresiones pueden contribuir a la normalización de actitudes hostiles, afectando la armonía social y generando un entorno de inseguridad en diferentes contextos.

Frente a esta problemática, la comunidad científica ha desarrollado diversos enfoques para la detección automática. [4] destacan que el discurso de odio sigue siendo problemático y desafiante, ya que tanto los seres humanos como los modelos de aprendizaje automático enfrentan dificultades para detectarlo debido a la complejidad y variedad de las categorías del discurso de odio. [5] propone el uso de algoritmos de Machine Learning (ML) para localizar discurso de odio en textos online en cuatro idiomas: inglés, español, italiano y portugués, los modelos que utilizaron fueron Naive Bayes (NB), Logistic Regression (LR) y Support Vector Machine (SVM), en donde los experimentos muestran que los mejores resultados alcanzan una precisión del 82,51 % y un valor F1 de alrededor del 83 %.

En la búsqueda de metodologías más avanzadas para la detección automática de discursos de odio, [6] investigaron técnicas que combinan representaciones específicas del lenguaje con el modelo BERT [7] (Bidirectional Encoder Representations from Transformers), este enfoque, basado en embeddings entrenados con datos etiquetados, mejora la

precisión en la clasificación y demuestra el potencial de las arquitecturas de aprendizaje profundo para superar las limitaciones de los modelos tradicionales.

En el análisis multilingüe, [8] abordaron la detección de discursos de odio, incluyendo el idioma español, mediante el uso del modelo denominado LASER con Regresión Logística, mBERT [9], y técnicas de traducción combinadas con BERT. Los experimentos realizados muestran que los mejores resultados en español se obtienen utilizando mBERT, alcanzando un valor F1 de 0.7329 en un escenario con entrenamiento completo.

En SemEval-2019 [10], se evaluaron modelos como Support Vector Machine con kernel lineal para la detección de odio hacia grupos vulnerables, logrando un macro F1 promedio de 0.73. Los modelos se entrenaron con representaciones de texto basadas en bag-of-words, bag-of-characters y embeddings orientados al sentimiento, demostrando su eficacia en la identificación de odio hacia inmigrantes y mujeres.

En el contexto de las redes sociales, [11] desarrollaron un detector automático de discurso de odio en español en Twitter mediante técnicas de aprendizaje supervisado, en donde se implementaron ocho modelos predictivos, entre los modelos de aprendizaje superficial, la Regresión Logística y el Multinomial Naïve Bayes (MNB) destacaron con un valor de F1 de 0.78.

[12] exploró diferentes algoritmos de aprendizaje automático para la clasificación de discurso de odio en Twitter, utilizando datos en indonesio, el conjunto de datos incluyó 4,002 tweets clasificados en categorías relacionadas con política, religión y etnicidad, entre los modelos evaluados, el algoritmo de MNB sin la técnica SMOTE obtuvo el mejor desempeño en términos de recall con un 93.2% y una precisión general del 71.2%, otros algoritmos probados, como SVM y Multi-Layer Perceptron (MLP), también mostraron buenos resultados, con un recall del 91.1% y una precisión del 83.4% al combinarse con SMOTE, respectivamente.

En la tarea compartida HOMO-MEX organizada en IberLEF2024, [13] propusieron detectar discurso de odio dirigido a la comunidad LGBT+ en publicaciones de Twitter y letras de canciones en español, para ello se emplearon modelos basados en Transformers como RoBERTa [14], XLM-RoBERTa [15] y BETO [16], logrando un F1 de hasta 91.43% en redes sociales, sin embargo, en el análisis de canciones, el desempeño fue menor, alcanzando solo un F1 de 57.62%, evidenciando la necesidad de enfoques específicos para textos estilísticamente complejos, como las letras de canciones.

Siguiendo con el uso de técnicas de ML, [17] analizan los efectos dañinos de las redes sociales, como el ciberacoso y la violencia contra grupos vulnerables y proponen dos clasificadores para detectar discursos de odio en Facebook, uno basado en SVM y otro en redes neuronales llamado Long Short-Term Memory (LSTM) [18], los resultados muestran que el modelo SVM alcanzó una precisión del 72.85% y un F1 de 0.594, mientras que el modelo LSTM obtuvo una precisión del 75.23% y un F1 de 0.657.

De igual manera, [19] propuso Multi-view SVM, un clasificador de discursos de odio que aplica SVM apilada de múltiples vistas. En donde cada tipo de característica se procesa con su propio clasificador SVM lineal, usando una constante de regularización de 0.1, creando un clasificador específico para cada vista de características.

Para mejorar la eficacia de estos clasificadores, la investigación se ha centrado también en las técnicas de extracción de características. [20], emplearon características basadas en n-gramas, específicamente unigramas y bigramas, para representar el texto en forma de conjuntos secuenciales de palabras extraídas de cada tweet, con el fin de optimizar el rendimiento de estas características. Este enfoque ha demostrado ser particularmente efectivo, como lo confirma [21] que utilizó el método Term Frequency-Inverse Document Frequency (TF-IDF) [22] como técnica principal de extracción de características en un sistema diseñado para clasificar textos provenientes de Twitter, permitiendo identificar discursos de odio al asignar mayor peso a los términos más relevantes dentro de cada documento, describiendo cómo este enfoque, en combinación con la Regresión Logística multinomial, resulta particularmente útil en la detección de patrones lingüísticos asociados a discursos negativos, destacándose como una herramienta clave para analizar grandes volúmenes de datos textuales, especialmente en contextos de redes sociales.

La efectividad de estas técnicas ha sido respaldada por diversos estudios comparativos. [23] evaluaron enfoques como n-gramas con TF-IDF, Word2Vec [24] y Doc2Vec [25], combinados con modelos como SVM, Random Forest y Naive Bayes, encontrando que la combinación de n-gramas con TF-IDF y SVM obtuvo el mejor desempeño, alcanzando un 79% de precisión. Mientras que [26] reportó una precisión del 88.77%, acompañada de un valor de recall de 88.77% y una puntuación F1 de 87.81% al emplear estas características en combinación con el algoritmo SVM.

Basado en estos estudios, el enfoque TF-IDF demuestra ser una herramienta valiosa para identificar patrones significativos en textos, consolidándose como una técnica ampliamente utilizada en investigaciones relacionadas con el análisis de datos textuales.

Los modelos ensamblados han emergido como una aproximación prometedora, como demuestra [27] en donde desarrollaron un sistema que combina tres clasificadores de ML: SVM, Logistic Regression y Random Forest, utilizando estrategias de votación para aprovechar las fortalezas individuales de cada clasificador y construir un sistema robusto.

La integración de múltiples enfoques en un modelo puede ofrecer ventajas significativas en términos de precisión y robustez, al combinar diferentes perspectivas para abordar un problema, este enfoque permite aprovechar las fortalezas individuales de cada componente, lo que podría resultar en predicciones más confiables. Asimismo, incorporar estrategias que reduzcan posibles sesgos podría contribuir a mejorar la capacidad del sistema para generalizar frente a diversas fuentes de datos.

La presente investigación experimenta con diferentes modelos tradicionales de clasificación de textos, incluyendo enfoques estadísticos y basados en redes neuronales, utilizando características fraseológicas y representaciones de n-gramas de TF-IDF para evaluar su efectividad en la detección automática del discurso de odio en textos. Este enfoque experimental tiene como objetivo explorar el rendimiento de técnicas ya establecidas, no con la intención de proponer mejoras a los modelos existentes, sino de analizar su aplicabilidad y desempeño en escenarios específicos de clasificación de discursos de odio. En particular, se busca comprender cómo estos métodos responden a la complejidad lingüística y los matices propios del discurso de odio, con el fin de aportar evidencia empírica que contribuya al desarrollo de herramientas más sólidas para este campo de estudio.

II. MÉTODO

Para el desarrollo de esta investigación se realizaron pruebas experimentales utilizando diversos modelos de clasificación y características específicas, con el objetivo de evaluar su rendimiento y efectividad en la tarea de clasificación de discursos de odio en español. En la Fig. 1 se muestra el flujo de trabajo seguido durante los experimentos.

Fig. 1 Esquema propuesto para la investigación.

El modelo propuesto sigue un enfoque secuencial para la detección de discursos de odio en textos, la primera etapa consiste en la selección del dataset a utilizar. Se realiza el preprocesamiento y la extracción de características, las cuales son fundamentales para el entrenamiento de los modelos, la extracción de características permite transformar estos textos en representaciones numéricas que los modelos pueden procesar.

Posteriormente, se procede a realizar el entrenamiento de los diferentes modelos de clasificación seleccionados. Durante esta fase, cada modelo aprende a identificar patrones en los datos procesados que le permiten distinguir entre textos que contienen discurso de odio y aquellos que no.

Finalizado el entrenamiento, se procede a la fase de predicción con los modelos entrenados. Por último, se compararon los resultados obtenidos del mejor modelo con otros modelos entrenados reconocidos para la detección de discurso de odio en textos.

A. *Dataset utilizado*

El dataset utilizado para el desarrollo de esta investigación fue OffendEs [28] un conjunto de datos diseñado específicamente para abordar la problemática del discurso de odio en español y que constituyó una parte fundamental de la competencia MeOffendES [29] sobre detección de contenido ofensivo en español en IberLEF 2021[1]. La tarea consistió en determinar si un texto contenía discurso de odio o no.

El dataset[2] está dividido en tres partes: entrenamiento, validación y prueba. La información contenida en el dataset proviene de diversas plataformas digitales; entre su contenido se incluyeron tweets de Twitter, comentarios de YouTube y publicaciones de Instagram. Cada dataset estuvo etiquetado con identificadores únicos (id), el texto del comentario, el nombre del influencer, su género, el dominio de origen del texto y la etiqueta correspondiente al tipo de odio contenido.

En la Tabla 1 se muestra el número de instancias del dataset.

TABLA 1
INSTANCIAS DEL DATASET

Entrenamiento	16.710
Validación	100
Pruebas	13.606

En la Fig. 2 se aprecia una muestra del dataset de entrenamiento.

	id	comment	influencer	influencer_gender	media	label
0	52564	En vez de la magia de mi melena, la magia de m...	dalas	man	instagram	NO
1	32984	A ver, los milenials y la gente normal necesit...	soyunapringada	woman	youtube	NO
2	58447	Me encanta todo el contenido que haces se nota...	wildhater	man	instagram	NO
3	10341	a Laura sige así que vales mucho más que 10 o ...	lauraescane	woman	youtube	NO
4	53087	Y si no mes gusta Dalas, que hacen aquí,lárgue...	dalas	man	instagram	NO
...	...	...	...	...	...	...
16705	57470	Hijo de tu puta madre estoy mamadisimo 😎	dalas	man	instagram	OFP
16706	35	yo que hace 4 años lo veía, ahora me doy cuent...	dalas	man	twitter	OFP
16707	18564	Esta re blanco el wismi	wismichu	man	youtube	OFP
16708	46485	algo que no veo en esa botella rosada es que s...	windygirk	woman	youtube	OFP
16709	47965	Ballena	soyunapringada	woman	youtube	OFP

16710 rows × 6 columns

Fig. 2 Contenido del dataset de entrenamiento.

B. *Preprocesamiento*

El dataset original se encuentra etiquetado de cuatro formas: OFP (Ofende a una persona), OFG (Ofende a un grupo), NOE (No odio explícito) y NO (No odio), de las cuales dos correspondían a texto ofensivo y dos a no ofensivo. Para efectos de la presente investigación, se dejaron solo dos etiquetas, considerando si el texto es ofensivo o no, con el propósito de realizar un entrenamiento de tipo clasificación binaria. Fue fundamental transformar los valores de la etiqueta objetivo a una representación que el modelo pudiera interpretar.

[1] https://sites.google.com/view/iberlef2021/home

[2] https://huggingface.co/datasets/SINAI/OffendES

Por lo tanto, dichos valores fueron codificados, asignando 0 a los textos que no contenían odio y 1 en caso de que el texto representara odio, lo que permitió la correcta clasificación binaria de los datos.

Se eliminaron elementos como, hashtags, enlaces, espacios en blanco y emoticones, además, los textos se convierten a minúscula para mantener un formato uniforme. Finalmente, se lematiza los textos utilizando SpaCy [30] con su librería es_core_news_sm, optimizado para el idioma español, con el fin de agrupar variantes morfológicas de una palabra bajo una única forma, para mejorar el análisis del texto, ejemplo corriendo – correr. La Fig. 3 refleja los resultados del preprocesamiento.

	id	comment	label	texto_limpio
0	27341	Me encanta el videooo porciento aidii he subid...	0	yo encantar el videooo porciento aidii haber s...
1	21310	Ropa cara?veo dulceida shop, Zara.. y de todas...	0	ropa cara?veo dulceida shop , zar .. y de todo...
2	23809	Y la perra seguia y seguia.jpg :v	1	y el perra seguia y seguia.jpg : v
3	14532	Malditas drogas	0	maldita droga
4	51651	perdona el spam , es la primera vez que trato ...	0	perdonar el spam , ser el primero vez que trat...
...	...	...	...	...
95	58740	Me alegra dalas . Soy nuevo seguidor y me aleg...	0	yo alegrar dalas . ser nuevo seguidor y yo ale...
96	6796	En resumen Dalas le callo la puta boca.	0	en resumen dala él callar el puta boco .
97	53415	Genial!! Que te lleven a Alcaser, a ver las Ca...	0	genial ! ! que tú llevar a alcaser , a ver el ...
98	49449	Hola dalas quiero que sepáis que tienes mi tot...	0	holar da él querer que sepáis que tener mi tot...
99	45149	Por que crees que hablan mal de ti 1 haces muc...	0	por que crees que hablar mal de tú 1 hacer muc...

100 rows × 4 columns

Fig. 3 Dataset de entrenamiento texto limpio y lematizado

C. *Extracción de características*

1) Fraseológicas

Con el fin de identificar patrones lingüísticos característicos del discurso de odio, se realizó un análisis exhaustivo de las características fraseológicas de los textos pertenecientes al conjunto de datos. Durante este análisis, se extrajeron diversas métricas claves que permitieron una evaluación detallada del contenido textual, como la longitud promedio de las palabras, la diversidad léxica y la variación en la estructura de las oraciones, En la Tabla 2 se aprecian las características fraseológicas extraídas.

TABLA 2
CARACTERÍSTICAS FRASEOLÓGICAS

Mean Word Length	Longitud media de palabras
Lexical Diversity	Diversidad léxica
Standar Deviation Sentence Length	Desviación estándar de la longitud de las oraciones
Mean Paragraph Length	Longitud media de los párrafos
Mean Sentence Length	Longitud media de las oraciones
Document Lenght	Longitud total del documento
Words per text	Número de palabras por texto analizado
Sentence per text	Número de oraciones por texto analizado
Mean Difference Sentence Length	Diferencia promedio entre las longitudes de las oraciones

Estas métricas se representaron numéricamente, proporcionando información sobre los aspectos lingüísticos predominantes en los textos. Los valores obtenidos a partir de estas métricas ayudaron a detectar patrones recurrentes y distintivos que son típicos del discurso de odio, como el uso de un vocabulario limitado pero cargado de connotaciones agresivas, o la repetición de ciertas estructuras sintácticas con el fin de intensificar el impacto del mensaje. En la Fig. 4 se observa una muestra de las características fraseológicas extraídas.

	id	MeanWordLen	LexicalDiversity	MeanSentenceLen	StdevSentenceLen	MeanParagraphLen	DocumentLen	WordsPerText	SentencesPerText
0	52564	3.000000	64.285714	15.0	0.0	15.0	57	14	1
1	32984	4.304348	91.304348	12.5	7.5	25.0	124	23	2
2	58447	3.812500	68.750000	16.5	6.5	33.0	154	32	2
3	10341	3.210526	84.210526	19.0	0.0	19.0	79	19	1
4	53087	4.114286	65.714286	124.0	0.0	124.0	574	105	1
...	...	...	...	...	...	...	...	...	...
16705	57470	4.571429	100.000000	7.0	0.0	7.0	38	7	1
16706	35	3.375000	87.500000	17.0	0.0	17.0	71	16	1
16707	18564	3.800000	100.000000	5.0	0.0	5.0	23	5	1
16708	46485	4.217391	66.666667	75.0	0.0	75.0	371	69	1
16709	47965	6.000000	100.000000	1.0	0.0	1.0	6	1	1

16710 rows × 10 columns

Fig. 4 Características fraseológicas extraídas

2) N-gramas para TF-IDF

Para la generación de los valores TF-IDF, se empleó un enfoque que considera tanto palabras individuales como diferentes combinaciones de palabras para la tarea en cuestión. Para ello, se codificó un tokenizador que segmentó el texto en unigramas, bigramas y trigramas, permitiendo identificar relaciones significativas entre las palabras. En la Tabla 3 se puede apreciar las formas en que los n-gramas agrupan el texto de la oración “esto es un artículo”.

TABLA 3
EJEMPLO DE SEPARACIÓN DE N-GRAMAS

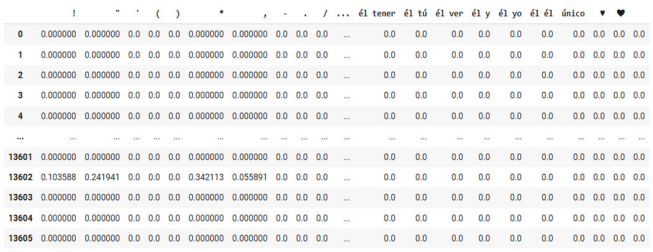
n-grama	**Ejemplo**
Texto ejemplo	“esto es un artículo”
unigrama	“esto”, “es”, “un”, “artículo”
bigrama	“esto es”, “es un”, “un artículo”
trigrama	“esto es un”, “es un artículo”

El proceso de tokenización y segmentación permitió preparar el texto para su posterior vectorización, asegurando que las combinaciones de palabras fueran representativas del contexto en el que aparecían. Este enfoque facilitó no solo la identificación de patrones léxicos en el texto, sino también la reducción de ruido en los datos, descartando combinaciones irrelevantes o redundantes.

Posteriormente, se calculan los valores de TF-IDF haciendo uso de tfidfVectorizer de Scikit-Learn [31], para cada una de las características extraídas, evaluando la importancia de las palabras y combinaciones dentro del conjunto de datos e identificando las características más relevantes para el análisis.

Para efectos de experimentación y evaluación de mejores resultados, se extrajeron valores de TF-IDF de 500, 1000, 1500 y 2000 características, con el propósito de determinar con qué

valor se obtienen los mejores resultados. En la Fig. 5 se observan las características TF-IDF de los textos.



	!	"	'	(	)	*	,	-	.	/	...	él tener	él tú	él ver	él y	él yo	él él	único	♥	❤
0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13601	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13602	0.103588	0.241941	0.0	0.0	0.0	0.342113	0.055891	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13603	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13604	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13605	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

13606 rows × 500 columns

Fig. 5 Total de características TF-IDF extraídas: 500

D. *Modelos de clasificación utilizados*

Con el propósito de este estudio, se utilizaron distintos modelos de clasificación, listados en la Tabla 4.

TABLA 4
MODELOS UTILIZADOS

Modelos de clasificación
Random Forest (RF)
K-Nearest Neighbors (KNN)
Logistic Regression (LR)
LinearSVC (LSCV)
Decision Tree (DT)
Multinomial Naive Bayes (MNB)
Multi-Layer Perceptron (MLP)
eXtreme Gradient Boost (XGB)
Voting Classifier (VC)
Stacking Classifier (SC)

Cada modelo seleccionado se entrenó con los datos de entrenamiento y se ajustaron sus parámetros a valores predefinidos recomendados para análisis de texto. Dichos modelos se utilizaron por su buen rendimiento demostrado en tareas similares de clasificación de textos [32].

Voting Classifier y Stacking Classifier fueron las técnicas de ensamblado de modelos utilizadas para repotenciar unificando las características de los modelos de clasificación descritos.

E. *Entrenamiento*

Para llevar a cabo un correcto entrenamiento de los modelos, fue necesario balancear el conjunto de datos, implementando técnicas como Over Sampling y SMOTE, siendo esta última la que mostró mejores resultados al lograr un mejor balance de las clases y, por ende, un desempeño superior del modelo. Tras el balanceo de los datos, se procedió con el escalado de las características utilizando Min-Max scaler, lo cual permitió ajustar las características numéricas a una escala uniforme en el rango [0,1], garantizando que todas las variables contribuyeran de manera equitativa al modelo, evitando así el dominio de características con valores mayores, optimizando así la eficacia de los modelos de clasificación.

Como se mencionó anteriormente, se realizaron pruebas con diferentes dimensiones de características TF-IDF de 500, 1000, 1500 y 2000, determinando que al emplear 2000 características se obtuvo un desempeño superior en términos de precisión y consistencia, finalmente, se procedió a guardar los modelos.

F. *Predicción y evaluación*

La predicción se realizó sobre el conjunto de datos de pruebas, del cual también se extrajeron las características fraseológicas y n-gramas de TF-IDF. Es de notar, este proceso se hizo con todos los modelos de clasificación y ensamblado mencionados, obteniendo el reporte de predicción de todos ellos. Sobre dichas predicciones, se generó la matriz de confusión correspondiente al modelo con mejor rendimiento. Es importante anotar que antes de esta fase, se realizó una evaluación previa con el dataset de validación para verificar el rendimiento de los modelos antes de proceder con el dataset de prueba.

G. *Comparación con modelos preentrenados*

Como parte del análisis, se llevó a cabo una comparación entre el desempeño del modelo propuesto y el de otros modelos preentrenados reconocidos, como Pysentimiento [33] y Piuba [34]. Esta comparación fue crucial para evaluar el desempeño relativo de los modelos entrenados en comparación con modelos establecidos en la literatura, proporcionando una referencia sobre cómo se comportan los modelos en un contexto similar.

Pysentimiento es una herramienta diseñada para facilitar el análisis de sentimientos y opiniones en investigaciones relacionadas con redes sociales. Basado en la biblioteca HuggingFace, proporciona una API sencilla que permite a los investigadores utilizar modelos de última generación para el análisis de sentimientos y otras tareas de PLN. Actualmente, admite los idiomas español, inglés, italiano y portugués, lo que lo posiciona como una opción versátil para estudios multilingües.

```
hate_speech_analyzer = create_analyzer(task="hate_speech", lang="es")
```

Fig. 6 Instancia del modelo Pysentimiento

Piuba es un modelo diseñado específicamente para la detección de comentarios de incitación al odio en artículos periodísticos. Basado en BERT, un modelo preentrenado en español ampliamente reconocido en el campo del PLN, Piuba aborda el desafío de la clasificación multietiqueta, asignando a cada entrada una etiqueta correspondiente a los distintos grupos considerados en el análisis. Piuba se implementa de la siguiente manera:

```
tokenizer = AutoTokenizer.from_pretrained("piuba-bigdata/beto-contextualized-hate-speech")
model = AutoModelForSequenceClassification.from_pretrained("piuba-bigdata/beto-contextualized-hate-speech")
```

Fig. 7 Instancia del modelo Piuba

III. RESULTADOS

La propuesta para detectar la presencia de discursos de odio en textos mediante el uso de modelos de clasificación con características fraseológicas y frecuencias de n-gramas de TF-IDF presentó los siguientes resultados:

Se experimentó con diversas cantidades de TF-IDF para evaluar si a mayor o menor cantidad de estas características se obtienen mejores resultados, como se muestra en Tabla 4.

TABLA 4
PREDICCIÓN MACRO F1 SEGÚN CANTIDAD TF-IDF

Modelo	500	1000	1500	2000
RF	0.662071	0.667812	0.676742	0.675693
KNN	0.296207	0.363717	0.293970	0.709679
LR	0.666711	0.694104	0.705002	0.304688
LSCV	0.661798	0.688969	0.695663	0.698878
DT	0.642957	0.661993	0.662910	0.669197
MNB	0.650977	0.679982	0.683344	0.687830
MLP	0.647058	0.680712	0.684080	0.702106
XGB	0.682697	0.707497	0.720872	0.721816
VC	0.706557	0.726313	0.733924	**0.742261**
SC	0.599074	0.616679	0.609363	0.617670

Se observó que al utilizar 2000 características TF-IDF, se obtuvieron mejores resultados, por lo que nuestras predicciones finales se harán en base a esta cantidad. La Tabla 5 muestra los resultados obtenidos para la predicción final.

TABLA 5
MÉTRICAS DE EVALUACIÓN DE LA PREDICCIÓN CON 2000 TF-IDF

Modelo	Accuracy	Precision	Recall	Macro F1
RF	0.842569	0.759954	0.646957	0.675693
KNN	0.305233	0.533679	0.529636	0.709679
LR	0.799500	0.693037	0.741524	0.304688
LSCV	0.797663	0.685166	0.721547	0.698878
DT	0.776055	0.657344	0.690609	0.669197
MNB	0.767603	0.672705	0.741194	0.687830
MLP	0.848743	0.767611	0.673222	0.702106
XGB	0.852933	0.770560	0.695551	0.721816
VC	0.846024	0.747355	0.737587	**0.742261**
SC	0.841394	0.823492	0.596177	0.617670

Se aprecia que el modelo ensamblado Voting Classifier, fue el que obtuvo mejores resultados con un macro F1 de 0.742261. Fig. 8 muestra la matriz de confusión correspondiente al mejor modelo Voting Classifier, donde se visualiza la distribución de aciertos y errores en las predicciones realizadas.

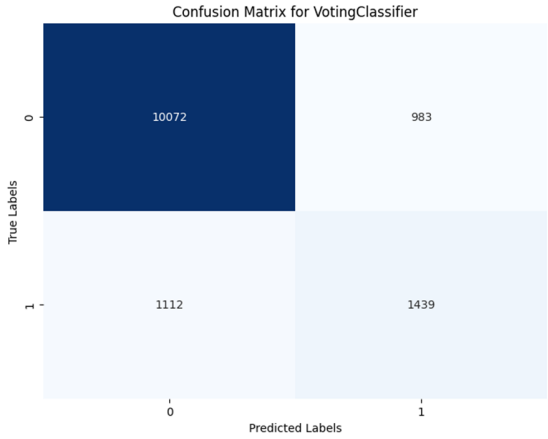


Fig. 8 *Matriz de confusión de la predicción*

El modelo registró 983 falsos positivos y 1,112 falsos negativos, mientras que alcanzó 10,072 verdaderos negativos y 1,439 verdaderos positivos. Estos valores reflejan un buen desempeño en la identificación de textos que no representan odio, aunque evidencia dificultades al clasificar correctamente los textos de odio.

A continuación, con el fin de comparar el desempeño del modelo Voting Classifier, se presentan las predicciones generadas por modelos reconocidos en la detección de discurso de odio, como Pysentimiento y Piuba, Tabla 6.

TABLA 6
MÉTRICAS DE EVALUACIÓN DE LOS MODELOS

Modelo	macro F1
Nuestros resultados	**0.742261**
Pysentimiento	0.685661
Piuba	0.629883

Se aprecia que Voting Classifier, como mejor modelo de la presente investigación, obtuvo mejores resultados que los modelos entrenados Piuba y Pysentimiento para la tarea de detección de odio en textos.

IV. Discusión

El análisis de los resultados obtenidos muestra que el desempeño de los modelos de clasificación mejora significativamente al incrementar la cantidad de características TF-IDF. Al utilizar 2000 características, el modelo Voting Classifier alcanzó un accuracy del 84%, reflejando un buen rendimiento general en la clasificación de textos entre odio y no odio. Este valor elevado de accuracy demuestra la capacidad del modelo para clasificar correctamente una proporción considerable de los textos, principalmente debido al predominio de la clase mayoritaria en el conjunto de datos. Sin embargo, el macro F1 del modelo, con un valor de 0.742, que es más bajo en comparación con el accuracy. Esto evidencia que el desempeño del modelo no es uniforme entre las clases, ya que el macro F1 toma en cuenta el balance entre precisión y recall

de ambas clases, penalizando el rendimiento en la clase minoritaria.

La interpretación de la matriz de confusión del mejor modelo, el ensamblado Voting Classifier, revela que, aunque logró identificar correctamente 1,439 casos como verdaderos positivos entre los 2,551 textos etiquetados como odio, también clasificó erróneamente 1,112 casos como falsos negativos. Por otro lado, para la clase de no odio, el modelo demostró una alta precisión, identificando correctamente 10,072 casos como verdaderos negativos frente a 983 falsos positivos. Este análisis sugiere que, a pesar del buen rendimiento global, existe una oportunidad de mejorar la identificación de textos de odio, especialmente reduciendo la proporción de falsos negativos y positivos.

Aunque el modelo mostró un buen rendimiento general, aún existe margen para mejorar la precisión en la identificación de textos de odio, especialmente en la reducción de los falsos positivos y negativos.

Voting Classifier alcanzó un macro F1 de 0.742, superando a Pysentimiento (0.686) y Piuba (0.630), demostrando un mejor equilibrio entre precisión y sensibilidad. Su desempeño superior se atribuye a la integración de características fraseológicas y n-gramas de TF-IDF, resaltando la importancia de enfoques personalizados sobre modelos preentrenados en contextos específicos como la detección de discursos de odio.

Estos resultados refuerzan la importancia de combinar técnicas de clasificación avanzadas con una selección adecuada de características. En particular, la incorporación de características fraseológicas y n-gramas de TF-IDF demostró ser fundamental para mejorar la capacidad de los modelos en la detección de discursos de odio en textos.

V. Conclusiones

La propuesta basada en la combinación de características fraseológicas y n-gramas de TF-IDF, implementada en modelos de clasificación, destacó por su efectividad en la detección de discursos de odio. En particular, el modelo ensamblado Voting Classifier alcanzó un macro F1 de 0.742, superando significativamente a los modelos preentrenados como Pysentimiento (0.686) y Piuba (0.630). Estos resultados evidencian la capacidad del enfoque para equilibrar precisión y sensibilidad en las clases de odio y no odio, posicionándolo como una alternativa sólida frente a modelos generales.

El análisis mostró que el incremento a 2000 de las características TF-IDF fue determinante para mejorar el desempeño de los modelos, siendo el Voting Classifier el más destacado en esta configuración. Asimismo, la integración de características fraseológicas permitió un mejor reconocimiento de patrones lingüísticos complejos y una mayor contextualización en la detección de expresiones explícitas e implícitas de odio.

Como trabajo futuro, sería valioso explorar la integración de técnicas basadas en aprendizaje profundo, como redes neuronales preentrenadas específicas para el idioma y el dominio, combinadas con las características fraseológicas propuestas.

En conclusión, este enfoque combina de manera efectiva técnicas avanzadas de selección de características y métodos de ensamblado, logrando capturar tanto patrones específicos como contextuales del lenguaje. Esto resalta su relevancia como herramienta para el desarrollo de sistemas más precisos en la clasificación de discursos de odio, especialmente en escenarios complejos y de alta variabilidad lingüística.

Referencias

[1] R. A. Guirao, "Observatorio del Sistema Penal y los Derechos Humanos Universidad de Barcelona DISCURSO DEL ODIO, PROTECCIÓN DE MINORÍAS Y SOCIEDAD DEMOCRÁTICA HATE SPEECH, PROTECTION OF MINORITIES AND DEMOCRATIC SOCIETY," 2019.

[2] E. S. Asemah, E. P. Nwaoboli, and Q. T. Nwoko, "Textual Analysis of Select Social Media Hate Speech Messages against Clergymen in Nigeria," 2022.

[3] M. Subramanian, V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Alexandria Engineering Journal*, vol. 80, pp. 110–121, Oct. 2023, doi: 10.1016/J.AEJ.2023.08.038.

[4] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Jun. 01, 2022, *MDPI.* doi: 10.3390/info13060273.

[5] E. K. Shepherd Arévalo, "Online hate speech detection using Machine Learning," Sep. 16, 2022. Accessed: Nov. 10, 2024. [Online]. Available: https://hdl.handle.net/20.500.14352/3308

[6] H. Saleh, A. Alhothali, and K. Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," *Applied Artificial Intelligence*, vol. 37, no. 1, Nov. 2021, doi: 10.1080/08839514.2023.2166719.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Jan. 18, 2025. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[8] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep Learning Models for Multilingual Hate Speech Detection," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.06465

[9] H. Xu, B. Van Durme, and K. Murray, "BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 6663–6675, Sep. 2021, doi: 10.18653/v1/2021.emnlp-main.534.

[10] V. Basile *et al.*, "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, pp. 54–63, 2019, doi: 10.18653/V1/S19-2007.

[11] J. J. Amores *et al.*, "Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español," *Cuadernos.info*, no. 49, pp. 98–124, 2021, doi: 10.7764/CDI.49.27817.

[12] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/830/3/032006.

[13] H. Gómez-Adorno *et al.*, "Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population," in *Procesamiento del Lenguaje Natural*,

Sociedad Espanola para el Procesamiento del Lenguaje Natural, Sep. 2024, pp. 393–405. doi: 10.26342/2024-73-30.

[14] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, Accessed: Jan. 19, 2025. [Online]. Available: https://arxiv.org/abs/1907.11692v1

[15] Alexis Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale." Accessed: Jan. 19, 2025. [Online]. Available: https://pytext.readthedocs.io/en/master/xlm_r.html

[16] Cañete, José and Chaperon, Gabriel and Fuentes, J.-H. and K. Rodrigo and Ho, and J. Hojin and Pérez, "Spanish Pre-Trained BERT Model and Evaluation Data." Accessed: Jan. 19, 2025. [Online]. Available: https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[17] F. Del *et al.*, "Hate me, hate me not: Hate speech detection on Facebook," 2017. [Online]. Available: http://www.alexa.com/topsites

[18] H. Okut, "Deep Learning for Subtyping and Prediction of Diseases: Long-Short Term Memory," 200AD. [Online]. Available: www.intechopen.com

[19] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS One*, vol. 14, no. 8, p. e0221152, Aug. 2019, doi: 10.1371/JOURNAL.PONE.0221152.

[20] O. Oriola and E. Kotze, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[21] P. Sari, B. Ginting, B. Irawan, and C. Setianingsih, "Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method," *Proceedings - 2019 IEEE International Conference on Internet of Things and Intelligence System, IoTaIS 2019*, pp. 105–111, Nov. 2019.

[22] M. Artama, I. N. Sukajaya, and G. Indrawan, "Classification of official letters using TF-IDF method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1516/1/012001.

[23] S. Abro, S. Shaikh, Z. Ali, S. Khan, G. Mujtaba, and Z. H. Khand, "Automatic Hate Speech Detection using Machine Learning: A Comparative Study," 2020. [Online]. Available: www.ijacsa.thesai.org

[24] K. W. Church, "Word2Vec," *Nat Lang Eng*, vol. 23, no. 1, pp. 155–162, Jan. 2017, doi: 10.1017/S1351324916000334.

[25] H. Dauda Abubakar, M. Umar, and M. A. Bakale, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *Journal of Science & Technology*, vol. 4, no. 1, pp. 27–33, 2022, doi: 10.56471/slujst.v4i.266.

[26] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 5, no. 2, pp. 79–83, May 2023, doi: 10.21512/emacsjournal.v5i2.9978.

[27] F. Balouchzahi, H. Lakshmaiah Shashirekha, and G. Sidorov, "HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier," 2021. [Online]. Available: https://mangaloreuniversity.ac.in/dr-h-l-shashirekha

[28] F. M. P. Del Arco, A. Montejo-Ráez, L. A. Ureña-lópez, and M.-T. Martín-Valdivia, "OffendES: A New Corpus in Spanish for Offensive Language Research," 2021. Accessed: Dec. 10, 2024. [Online]. Available: https://aclanthology.org/2021.ranlp-1.123

[29] F. M. Plaza-Del-Arco *et al.*, "Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 183–194, Sep. 2021, doi: 10.26342/2021-67-16.

[30] A. Sharma, R. Aggarwal, and R. Alawadhi, "A Comparative Study of Text Summarization using Gensim, NLTK, Spacy, and Sumy Libraries," vol. 19, Apr. 2023, [Online]. Available: http://xisdxjxsu.asia

[31] Kunai Jain, "Sklearn In Python, SciKit Learn In Python | Analytics Vidhya." Accessed: Dec. 04, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/

[32] J. Ortiz-Zambrano, C. Espin-Riofrio, and A. Montejo-Ráez, "Transformers for Lexical Complexity Prediction in Spanish Language," *Procesamiento del Lenguaje Natural*, vol. 69, pp. 177–188, Sep. 2022, doi: 10.26342/2022-69-15.

[33] J. M. Pérez *et al.*, "pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks," *ArXiv*, p. arXiv:2106.09462, Jun. 2021, doi: 10.48550/ARXIV.2106.09462.

[34] Pérez *et al.*, "Assessing the impact of contextual information in hate speech detection." Accessed: Dec. 08, 2024. [Online]. Available: https://huggingface.co/piuba-bigdata/beto-contextualized-hate-speech/blob/main/README.md