

GLOBAL PROJECT

ALUMNA:

Leticia Cervieri Lores

GRUPO:

Grupo 8

PROGRAMA:

Postgrado en Inteligencia Artificial y Machine Learning (10ª promoción)

2023 - 2024

NOMBRE DEL PROYECTO:

Análisis del lenguaje tóxico en español: un enfoque basado en NLP

Contenido

RESUMEN.....	3
INTRODUCCIÓN.....	4
ESTADO DEL ARTE.....	7
1. Relevamiento de literatura sobre detección de lenguaje tóxico.....	7
2. Relevamiento de literatura sobre sesgo en NLP.....	11
3. Relevamiento de datasets disponibles.....	13
4. Relevamiento de recursos disponibles y modelos aplicables.....	14
OBJETIVOS.....	14
Objetivo general.....	15
Objetivos específicos.....	15
SOLUCIÓN PLANTEADA.....	16
METODOLOGÍA.....	17
DESARROLLO DE CADA ETAPA.....	18
Etapa 1: Recolección de datos y análisis inicial.....	18
Etapa 2: Preprocesamiento de datos.....	21
Etapa 3: Análisis exploratorio de datos (EDA).....	21
Etapa 4: Modelado y evaluación.....	22
4.1 Importación de librerías y preparación de datos.....	23
4.2 Modelos estadísticos de Machine Learning.....	24
4.2.1 Regresión Logística.....	24
4.2.2 Support Vector Machine (SVM).....	24
4.2.3 Random Forest.....	24
4.3 Modelos de Deep Learning con transformers.....	25
4.3.1 DistilBERT.....	25
4.4 Evaluación de modelos.....	26
Etapa 5: Análisis de resultados y conclusiones.....	26
EVALUACIÓN.....	27
1. Metodología de evaluación.....	27
1.1. Definición de métricas de evaluación.....	27
1.2. División en datos de entrenamiento y test.....	28
2. Implementación de los modelos.....	28
3. Análisis de resultados.....	29
3.1. Comparación de modelos.....	29
3.2. Evaluación por región.....	29
4. Resultados y reflexiones finales.....	29
RESULTADOS.....	30
1. Desempeño de los modelos.....	30
1.1. Logistic Regression.....	30
1.2. Support Vector Machine (SVM).....	31
1.3. Random Forest.....	32
1.4. DistilBERT.....	32
2. Conclusiones generales.....	33
3. Propuestas de mejora.....	34
CONCLUSIONES Y TRABAJOS FUTUROS.....	36
1. Relación con los objetivos.....	36
2. Interpretación de los resultados.....	36
3. Limitaciones y trabajos futuros.....	37
REFERENCIAS.....	39
ANEXOS.....	41

RESUMEN

Un análisis de la toxicidad online

El crecimiento exponencial de las redes sociales ha facilitado la difusión de información, pero también ha dado lugar a un aumento alarmante de la toxicidad en el discurso, particularmente en plataformas como Twitter (actualmente X). Este proyecto aborda el problema de los tweets ofensivos, cuestión que no afecta solo a la calidad de los contenidos publicados online, sino que tiene repercusiones también a nivel social, cultural y personal, afectando diferentes áreas como la salud mental, la representación de grupos o la cohesión social. A través de un análisis detallado de un dataset conformado por tweets en idioma español, se busca clasificar los mensajes como ofensivos o no ofensivos, centrándose en la influencia del idioma y la región de origen en la percepción de la toxicidad.

Para abordar este desafío, se ha desarrollado un enfoque desde el Procesamiento del Lenguaje Natural (NLP) en el que se combinan técnicas de Machine Learning estadístico (ML) y de Deep Learning (DL). Se emplearon para ello varios modelos diferentes: Regresión Logística, Random Forest y Support Vector Machine (SVM), para el aprendizaje automático más tradicional, y un modelo de transformers con BERT (DistilBERT), para el aprendizaje profundo con redes neuronales. El dataset utilizado está conformado por un conjunto de casi 30.000 tweets en idioma español, provenientes de la fusión de distintos datasets ya existentes y etiquetados manualmente, por lo que constituye una novedad frente a investigaciones previas. La solución propuesta no solo se centra en la detección de lenguaje ofensivo a través de una clasificación binaria (tóxico vs. no tóxico), sino que también evalúa el sesgo potencial asociado a la variedad del idioma de donde proviene el contenido (español de Latinoamérica vs. español de España) y cómo esto puede afectar a las predicciones de toxicidad.

Los resultados globales de los modelos implementados muestran que DistilBERT y SVM son los más efectivos para la detección de lenguaje tóxico, alcanzando una precisión y recall equilibrados en ambas clases. DistilBERT sobresale en la clase tóxica con un f1-score de 0.75, lo que indica su capacidad superior para identificar lenguaje ofensivo en comparación con los modelos tradicionales que mostraron más dificultades para detectar correctamente la clase tóxica. En cuanto a los resultados por región, los modelos tienden a funcionar mejor en Latinoamérica que en España, especialmente en la detección de tweets no tóxicos. En España, se observó una ligera disminución en la capacidad de los modelos para detectar tweets tóxicos, particularmente en DistilBERT, donde el recall de la clase tóxica fue más bajo.

Finalmente, se proponen estrategias y recomendaciones para futuros trabajos que incluyen la implementación de modelos más equitativos y entrenados con datasets que reflejen una mayor diversidad, el uso de modelos pre-entrenados más especializados o la exploración de características lingüísticas adicionales para mejorar el rendimiento del sistema en diversas regiones.

ADVERTENCIA DE CONTENIDO: Debido al tema de esta investigación, ciertos ejemplos pueden resultar ofensivos. Por este motivo, hemos minimizado al máximo la cantidad de muestras visibles en este informe.

INTRODUCCIÓN

Detectando la toxicidad en el idioma español

En el contexto de la creciente presencia de redes sociales y la abundancia de contenido generado por los usuarios, surge un problema crucial: la detección automática de lenguaje tóxico en plataformas como Twitter¹. Este problema es particularmente desafiante en este tipo de plataformas multilingües, donde el lenguaje, los dialectos y las expresiones varían considerablemente según la región y la cultura, así como en cualquier empresa u organización que opere con contenidos generados por usuarios de orígenes diversos que presentan multiplicidad de idiomas o variantes idiomáticas. En este trabajo, se propone un acercamiento a la detección de lenguaje tóxico en español, prestando especial atención a las diferencias entre las variedades lingüísticas de Latinoamérica y España.

Muchos modelos de procesamiento del lenguaje natural aplicados a la detección de lenguaje tóxico tienden a ser sesgados por región y/o idioma. Esto significa que un modelo entrenado con datos de una región podría no funcionar igual de bien en otra. En este caso, el sesgo regional de los tweets en español (entre Latinoamérica y España) es un aspecto crucial que afecta negativamente la capacidad de los modelos para detectar de manera justa y precisa los tweets tóxicos en ambas regiones.

El problema se ha identificado mediante la evaluación de varios modelos de clasificación de texto que, al aplicarse a los datos de diferentes regiones hispanohablantes, muestran un rendimiento significativamente diferente. En ciertos casos los modelos presentan una mayor precisión y recall en una determinada región, mientras que en la otra, la capacidad de detectar lenguaje tóxico disminuye considerablemente.

Históricamente, la detección de lenguaje tóxico ha sido abordada con una variedad de modelos tradicionales de Machine Learning y, más recientemente, con modelos avanzados de Deep Learning como BERT y sus variantes. Estos enfoques han incluido técnicas como:

- Modelos de Machine Learning tradicionales (por ejemplo, Logistic Regression, Support Vector Machine, Random Forest) que han sido usados con técnicas como TF-IDF para convertir texto en vectores numéricos y, a menudo, ajustados con estrategias de resampling para manejar datos desbalanceados.
- Modelos pre-entrenados de transformers, como BERT, que han demostrado ser efectivos en tareas de clasificación de texto debido a su capacidad para aprender representaciones complejas del lenguaje.

Sin embargo, estas soluciones tienden a ser globales y no siempre toman en cuenta las variaciones lingüísticas regionales, lo que ha llevado a resultados desiguales cuando los modelos se aplican a datos de diferentes regiones. En muchos casos, el sesgo por región no ha

¹ Aunque actualmente el nombre de la plataforma haya pasado a ser “X”, para este informe nos seguiremos refiriendo a ella como “Twitter” debido a que casi todos los contenidos analizados fueron generados previamente al cambio de denominación.

sido adecuadamente abordado, lo que ha afectado a la generalización de los modelos en diferentes contextos lingüísticos.

Se propone entonces una solución que combina enfoques tradicionales y avanzados de Machine Learning y Deep Learning para abordar de manera específica el sesgo regional en la detección de lenguaje tóxico en español. La solución planteada se basa en los siguientes elementos clave:

1. Ajustes específicos: Se sugieren técnicas de ajuste de pesos de clase para manejar el desbalance entre los tweets tóxicos y no tóxicos.
2. Uso de modelos pre-entrenados de Deep Learning: Se implementa DistilBERT, una versión ligera de BERT, que se ajusta a los datos en español y se combina con técnicas de Machine Learning estadístico tradicionales para comparar su rendimiento.
3. Evaluación diferenciada por región: En lugar de utilizar un modelo único para todos los tweets en español, se propone evaluar los modelos por separado para las regiones de Latinoamérica y España.

Esta solución busca abordar directamente el sesgo por región, un aspecto que ha sido pasado por alto en muchos enfoques anteriores. Al evaluar y ajustar los modelos por región, se busca crear un sistema más justo y eficaz para detectar lenguaje tóxico en las diversas variantes del español, lo que resulta esencial en un entorno multilingüe como el de las redes sociales.

El procedimiento seguido para implementar la solución incluye varios pasos:

1. Carga y preprocesamiento de datos: Se comenzó con la carga de un dataset preprocesado de tweets en español, asegurando que los textos estuvieran limpios y tokenizados.
2. División por región: Los datos fueron divididos en función de la región de origen de los tweets (Latinoamérica y España) para realizar análisis por separado.
3. Entrenamiento de modelos estadísticos de ML: Se implementaron modelos de Logistic Regression, SVM y Random Forest, utilizando representaciones TF-IDF de los textos y aplicando técnicas de balanceo de clases.
4. Entrenamiento de modelos de deep learning: Se implementó DistilBERT, ajustando el modelo pre-entrenado para detectar lenguaje tóxico en los datos en español.
5. Evaluación por región: Se evaluó cada modelo por separado en los subconjuntos de datos de Latinoamérica y España, comparando el rendimiento en ambas regiones para detectar posibles diferencias o sesgos.
6. Ajustes por región: Se realizaron ajustes en los pesos de clase para intentar mejorar el rendimiento en la clase con menor precisión, principalmente en España, donde el modelo mostró mayor dificultad en detectar tweets tóxicos.

Los resultados globales de los cuatro modelos implementados muestran que DistilBERT y SVM son los más efectivos para la detección de lenguaje tóxico, alcanzando una precisión y recall equilibrados en ambas clases. DistilBERT sobresale en la clase tóxica con un f1-score de 0.75, lo que indica su capacidad superior para identificar lenguaje ofensivo en comparación con los modelos tradicionales como Logistic Regression y Random Forest, que mostraron dificultades

para detectar correctamente la clase tóxica. Globalmente, los modelos lograron una precisión de entre 0.79 a 0.81, con AUC-ROC de hasta 0.85.

En cuanto a los resultados por región, los modelos tienden a funcionar mejor en Latinoamérica que en España, especialmente en la detección de tweets no tóxicos. Modelos como SVM y Random Forest alcanzaron un AUC-ROC cercano a 0.98 en Latinoamérica, lo que indica una excelente capacidad de discriminación entre clases. Sin embargo, en España, se observó una ligera disminución en la capacidad de los modelos para detectar tweets tóxicos, particularmente en DistilBERT, donde el recall de la clase tóxica fue más bajo. Estas diferencias sugieren que los factores lingüísticos y culturales regionales impactan en el rendimiento de los modelos, destacando la necesidad de realizar ajustes específicos por región.

Este documento se estructura en varias secciones. En la sección “Estado de arte” se presenta un resumen de la literatura existente y los enfoques anteriores sobre el tema, seguido de los “Objetivos”, tanto generales como específicos del proyecto. A continuación, se encuentra la “Solución planteada” con el detalle de los modelos utilizados y las técnicas específicas aplicadas para abordar el problema de sesgo por región. En el apartado de “Metodología” se describe el procedimiento detallado seguido para implementar la solución, incluyendo los pasos de recolección de datasets, preprocesamiento de datos, división por región, entrenamiento de modelos y evaluación. Posteriormente, la sección de “Evaluación” describe las métricas utilizadas y la metodología seguida para evaluar el rendimiento de los modelos y en “Resultados” se incluye un análisis exhaustivo del desempeño de cada uno, comparando su rendimiento. Finalmente, se presentan las “Conclusiones y trabajos futuros” donde se discuten los hallazgos del estudio y se sugieren recomendaciones sobre cómo mejorar la detección de lenguaje tóxico en el futuro, especialmente en contextos multilingües y multiculturales.

ESTADO DEL ARTE

Panorama actual en la detección de lenguaje tóxico: modelos, desafíos y oportunidades

El análisis de la toxicidad en el lenguaje online, especialmente en plataformas de redes sociales como Twitter, ha cobrado gran relevancia en los últimos años. Este fenómeno ha sido abordado desde múltiples perspectivas, empleando diversas técnicas NLP.

Al comenzar este estudio, se realizó una búsqueda en repositorios científicos y académicos generales (como Google Scholar, Arxiv, Semantic Scholar o Springer)² o específicos del área del NLP (ACL Anthology)³, así como en los sitios de competiciones relacionadas con NLP, como IBERLef, SemEval o TASS⁴, para detectar investigaciones, papers e informes relevantes en el área.

A continuación, se presentan los estudios, proyectos y recursos que hemos considerado más interesantes a la hora de contribuir en la comprensión y solución del problema de la toxicidad en el discurso en línea.

1. Relevamiento de literatura sobre detección de lenguaje tóxico

El crecimiento en la publicación de los contenidos generados por usuario (UGC) y en el uso de las redes sociales ha impulsado la investigación en el análisis de sentimientos y la detección automática de lenguaje tóxico, tareas críticas en el campo del NLP. En los últimos años, múltiples enfoques han sido propuestos para abordar este problema, y la investigación ha producido una variedad de métodos que van desde enfoques basados en léxicos hasta el uso de modelos de Deep Learning y modelos de lenguaje pre-entrenados.

La revisión de la literatura incluye diversas contribuciones que se pueden agrupar en tres categorías principales: enfoques basados en léxicos, modelos tradicionales de Machine Learning, y modelos avanzados de Deep Learning. Estos enfoques han sido evaluados en una amplia gama de conjuntos de datos que cubren diferentes idiomas, dominios y plataformas, desde comentarios en noticias hasta interacciones en redes sociales como Twitter o Reddit.

En Jahan & Oussalah (2023), siguiendo el método PRISMA, se hace una revisión de la literatura sobre el tema enfocándose sobre todo en el análisis de las diferentes definiciones del discurso de odio y sus conceptos relacionados (como ciberbullying, lenguaje ofensivo, discriminación o radicalización, tal como se muestra en la *Figura 1*), así como en el proceso general para llevar a cabo las aplicaciones de modelos de Machine Learning y Deep Learning que permitan detectar este tipo de discursos: recolección y preprocesamiento del dataset, feature engineering, entrenamiento de modelos y evaluación de resultados (*Figura 2*).

² <https://scholar.google.com/> , <https://arxiv.org/> , <https://link.springer.com/>,
<https://www.semanticscholar.org/>

³ <https://aclanthology.org/>

⁴ <http://sepln2023.sepln.org/iberlef/>, <https://semeval.github.io/>, <http://tass.sepln.org/2020/>

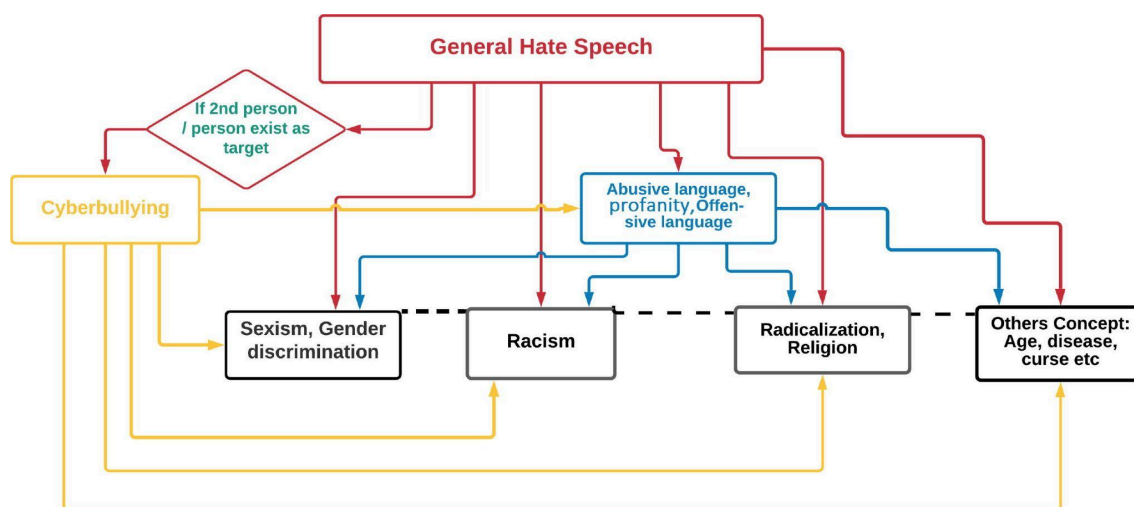


Figura 1. Diagrama relacional de conceptos en los discursos de odio. Tomado de Jahan & Oussalah (2023). <https://www.sciencedirect.com/science/article/pii/S0925231223003557#f0005>

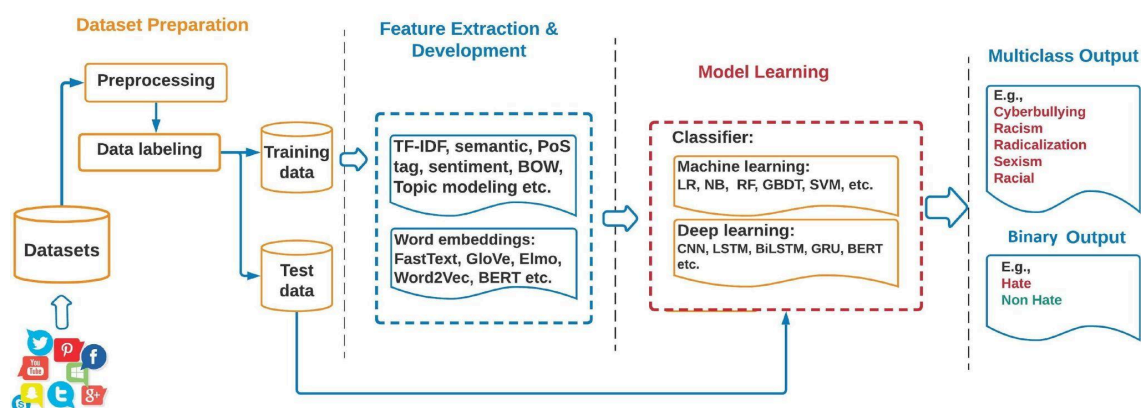


Figura 2. Pipeline general para la detección automática del discurso de odio. Tomado de Jahan & Oussalah (2023). <https://www.sciencedirect.com/science/article/pii/S0925231223003557#f0010>

Entre los resultados de este análisis de fuentes se encuentra que, entre el año 2000 y 2021, el 51% de los papers sobre el tema aplican sus modelos a datasets en idioma inglés, mientras que solo un 1% de las investigaciones fueron aplicadas al idioma español.

1.1 Enfoques basados en léxicos

Los enfoques basados en léxicos utilizan listas de palabras predefinidas para identificar comentarios tóxicos, hate speech o contenido ofensivo. Estas listas pueden incluir diccionarios específicos para palabras ofensivas, expresiones de odio o incluso léxicos para análisis de sentimientos. Ejemplos destacados incluyen:

- HurtLex⁵: Un léxico multilingüe diseñado para detectar lenguaje ofensivo a partir de una lista de categorías que abarcan insultos relacionados con etnicidad, religión, género, entre otros.
- MOL (Multilingual Offensive Language) y DALC (Dutch Abusive Language Corpus)⁶: Enfoques que combinan varios léxicos para mejorar la cobertura en diferentes dominios.
- Hatebase⁷: Un repositorio colaborativo que se actualiza con nuevas palabras y frases de hate speech en varios idiomas.
- NRC Emotion Lexicon⁸: Un enfoque léxico en inglés que mapea palabras a emociones, útil para identificar palabras con carga emocional negativa o que puedan considerarse ofensivas.

Algunos ejemplos de la aplicación de estos enfoques basados en lexicon pueden encontrarse en Liu (2020) y en Sahin et al. (2018), donde se muestra la aplicación de diccionarios predefinidos para comprobar la ocurrencia de determinadas palabras o patrones sintácticos en el texto a analizar, a través de métodos como la bolsa de palabras (BOW - Bag Of Words), los n-gramas, o la anotación de partes del discurso (POS - Part-of-Speech tagging).

Aunque estos modelos funcionan bien para determinados problemas de NLP (como por ejemplo la clasificación de correos basura en spam vs.no spam), no son capaces de explicar la semántica del lenguaje ya que tienden a ser poco efectivos en la detección de contexto, lo cual resulta fundamental para un problema de clasificación tan complejo como el de la detección del discurso de odio.

1.2 Modelos estadísticos de Machine Learning

Los enfoques tradicionales de ML estadístico aplicados a la clasificación de lenguaje tóxico incluyen algoritmos como Naïve Bayes, Support Vector Machine (SVM) y Random Forest, a menudo combinados con representaciones textuales como TF-IDF. Estos enfoques han sido exitosos en tareas de clasificación de texto, particularmente en contextos donde los conjuntos de datos son relativamente pequeños o donde no se dispone de modelos pre-entrenados específicos para el idioma.

Un ejemplo relevante en este sentido puede encontrarse en el trabajo de Koratana & Hu (2018) que compara la aplicación de un modelo de Logistic Regression con otros que utilizan redes neuronales convolucionales CNN y RNN con embeddings para detectar discurso de odio aplicado al dataset de la competición “Google Jigsaw Toxic Comment Classification Challenge”⁹, concluyendo que los modelos con embeddings proporcionan una mejora significativa en la métrica de accuracy.

⁵ <https://github.com/valeriobasile/hurtlex>

⁶ <https://github.com/franciellevargas/MOL>, <https://github.com/tommasoc80/DALC>

⁷ <https://hatebase.org/>

⁸ <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

⁹ <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

1.3 Modelos de Deep Learning y Transformers

En los últimos años, los modelos basados en DL han demostrado ser los más prometedores en la tarea de clasificación de lenguaje tóxico. Modelos como LSTM y, más recientemente, los modelos que utilizan transformers han sido entrenados para capturar las relaciones complejas entre las palabras que aparecen en el texto. En el ámbito del idioma español, los modelos basados en BERT y sus variantes han sido los más utilizados, entre los que destacan:

- BETO: Un modelo basado en BERT, pre-entrenado específicamente en grandes corpus de texto en español, ha mostrado buenos resultados en la clasificación de lenguaje tóxico en comentarios y tweets en español (Cañete et al., 2020).
- BERTIN: Una variante de RoBERTa entrenada en corpus español, enfocada en mejorar el rendimiento en la clasificación de texto en español (De la Rosa et al., 2022).

El uso de modelos multilingües, como mBERT y XLM-RoBERTa, también ha sido explorado en tareas de clasificación de texto multilingüe. Estos modelos, entrenados en más de 100 idiomas, tienen la ventaja de poder abordar simultáneamente varios idiomas y dialectos, como se muestra en el trabajo de Pires et al. (2019).

En la competición IberLEF de 2021, se evaluaron modelos como BETO y BERTIN en la tarea de detección de lenguaje tóxico en comentarios en español, obteniendo resultados superiores en comparación con los modelos tradicionales de Machine Learning. El trabajo de Plaza-del-Arco et al. (2021) en el corpus NECOS-TOX confirmó que los modelos basados en transformers son más efectivos para la detección de toxicidad en comentarios en noticias y redes sociales.

1.4 Clasificación con Large Language Models

Un enfoque más reciente incluye el uso de grandes modelos de lenguaje (LLM), como por ejemplo GPT-4, para clasificar comentarios tóxicos. Estos modelos han demostrado ser efectivos en configuraciones de clasificación de múltiples tareas, con capacidad de ajustar sus respuestas a diferentes instrucciones o prompts, lo que los convierte en una herramienta flexible para la clasificación de lenguaje tóxico en diferentes idiomas y dominios.

Así, en Hu & Zhang (2024) se analiza la detección de lenguaje por parte de grandes modelos de lenguaje al momento de rechazar prompts considerados ofensivos o maliciosos. Generalmente, lo que hacen estos modelos es tener una capa extra de fine tuning para la detección de este tipo de lenguaje, tanto en el input proveniente del prompt, como en el output generado por el LLM, lo que les genera costes adicionales y tiene alta tasa de errores, por lo que proponen usar lo que llaman MULI (Moderation Using LLM Introspection), un modelo de regresión logística que examina los logits del primer token del output para calcular la probabilidad de rechazar la generación de una respuesta al considerar que el input no es adecuado. Así, cuando hay respuestas cuyo primer token tiene altas probabilidades de ser “*Sorry, ...*” o “*Cannot...*”, se considera que el input tiene altas probabilidades de ser inadecuado. Ver *Figura 3*.

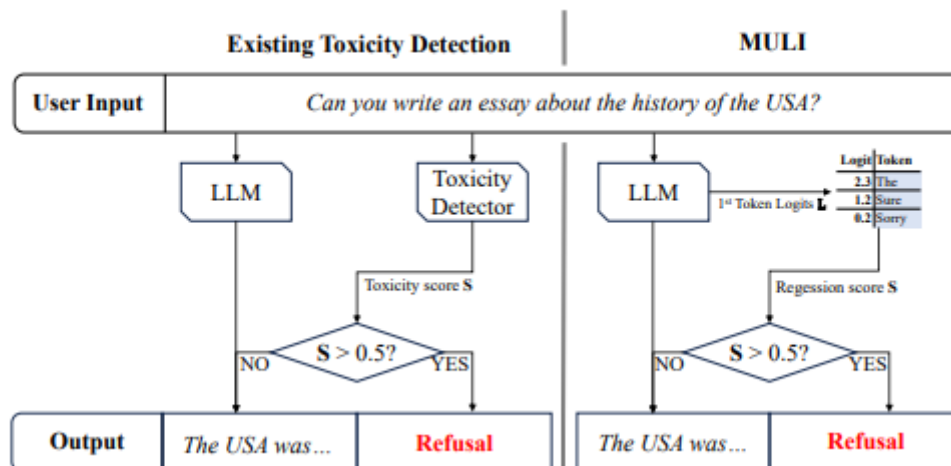


Figura 3. Pipeline de MULI. Tomado de Hu & Zhang (2024). <https://arxiv.org/pdf/2405.18822>

Relacionado con esto, en Zhang et al. (2023) se propone un método llamado Decision-Tree-of-Thought (DTot) para la detección de contenido inadecuado por parte de LLM aplicado a un dataset interno de Amazon, y en Guo et al. (2024) se analiza la importancia del contexto y de la definición del prompt para la detección de lenguaje ofensivo por parte de LLM (distinguiendo por ejemplo entre prompts generales, prompts con definición de Hate Speech, prompts few-shot learning y prompts chain-of-thought).

2. Relevamiento de literatura sobre sesgo en NLP

El análisis del sesgo en los sistemas de NLP ha sido un tema de creciente interés, especialmente en tareas como la detección de lenguaje tóxico, donde las diferencias lingüísticas y culturales pueden afectar la precisión de los modelos.

En Jurafski & Martin (2024), uno de los principales manuales para el estudio y la aplicación de NLP, se comenta que algunos clasificadores de toxicidad ampliamente utilizados identifican de forma incorrecta como tóxicas frases que simplemente contienen menciones a identidades como mujeres, personas ciegas u homosexuales o que utilizan formas lingüísticas características de variedades como el inglés vernáculo afroamericano, lo que puede llevar a un problema de silenciamiento en el discurso particular de o sobre estos grupos. Los modelos pueden entonces contener sesgos tanto en los datos de entrenamiento (al replicar o incluso intensificar los existentes en la sociedad), como en las propias etiquetas generadas para el análisis (debido a sesgos provenientes de los etiquetadores humanos), por los recursos utilizados (como léxicos o los embeddings de modelos pre-entrenados).

Siguiendo en esta línea, Mostafazadeh Davani et al. (2023) estudia el impacto de los estereotipos sociales en el comportamiento de los anotadores, en los conjuntos de datos

anotados y en los propios modelos de clasificación del discurso de odio, así como de las diferencias a nivel comunicativo de los sistemas lingüísticos.

Así también, la falta de modelos, datos de entrenamiento y de una evaluación eficaz a nivel multilingüe dificulta el desarrollo de modelos de detección de discursos de odio de mayor calidad para otros idiomas más allá del inglés. Como consecuencia, miles de millones de personas que no hablan inglés recibirán menos protección contra el odio en línea, ya que incluso las plataformas de redes sociales más grandes tienen claras brechas lingüísticas en su moderación de contenidos (Röttger et al. 2022).

2.1 Sistemas multilingües para la detección del lenguaje tóxico

En cuanto a la relevancia del idioma en el rendimiento de los modelos, encontramos sobre todo investigaciones y resultados provenientes de diversas competiciones para la aplicación de modelos de clasificación de lenguaje tóxico en contextos multilingües. Así, Mandl et al. (2021) y Mnassri et al. 2024 presentan los resultados de sus modelos implementados para lenguas indo-europeas como el hindi, el alemán, el bangla y el inglés, en el marco de la competición HASOC (Hate Speech and Offensive Content Identification), donde encontramos por ejemplo enfoques innovadores que combinan redes GAN (Generative Adversarial Networks) con modelos de lenguajes pre-entrenados como mBERT and XLM-RoBERTa, de manera de solventar el problema de la falta de datos anotados en diversos idiomas más allá del inglés.

Así también, el Multilingual HateCheck (MHC) se presenta como una herramienta diagnóstica conformada por una serie de tests funcionales para evaluar el funcionamiento de los modelos en la detección del lenguaje tóxico, aplicable a nivel de varios idiomas como árabe, francés, alemán, italiano, polaco, hindi, chino, portugués y español (Röttger et al. 2022).

En el trabajo de Das et al. (2024), por su parte, se propone un método basado en transformers para detectar hate speech en redes sociales como Twitter, Facebook, WhatsApp, Instagram, etc. El modelo propuesto se presenta como independiente del idioma, testeado de manera efectiva por ejemplo para datos en italiano, inglés, alemán y bengalí.

2.2 Detección del lenguaje tóxico en español

En cuanto a modelos de detección de lenguaje tóxico aplicados específicamente para el idioma español, encontramos ejemplos como Plaza-del-Arco et al. (2021), con la aplicación de modelos de Machine Learning (tanto SVM, como Logistic Regression) y Deep Learning (LSTM, CNN and Bi-LSTM) y modelos pre-entrenados (BERT, BETO) a datasets anotados en Hate Speech para idioma español, como es el caso del dataset HaterNet.

En Arango Monnar et al. (2022) se presenta la aplicación de modelos sobre un dataset anotado para tweets chilenos, destacando la importancia de llevar a cabo este tipo de investigaciones en otros idiomas más allá del inglés. Tal como comentan los autores, el fenómeno del discurso de odio depende estrechamente del contexto sociocultural al que pertenece, con lo que las características específicas del español hablado en diferentes países hacen que los modelos actuales sean poco generalizables y tengan bajo rendimiento cuando son entrenados en una

variedad lingüística y aplicados a otra (en su caso, referido al español de Chile frente al de España).

En cuanto a las competiciones referidas a datos para idioma español, podemos destacar IberEval, donde para la edición de 2018 se presentan las tareas AMI (Automatic Misogyny Identification), con el objetivo de detectar lenguaje sexista en español e inglés, tanto en clasificación binaria (comentario misógino vs. no misógino), como en la identificación del tipo de comportamiento (estereotipo, descrédito, violencia, acoso, etc.) y target (genérico vs. individual, es decir, ataques dirigidos a un grupo o a una persona en concreto), y MEX-A3T para el perfilado de autor y la detección de agresividad en tweets mexicanos y españoles, con el objetivo de impulsar la investigación para el tratamiento de variedades del español que presentan rasgos culturales significativamente diferentes a los del español peninsular. (Fersini et al. 2018, y Alvarez-Carmona et al. 2018).

Otras iniciativas en esta línea son las de SemEval (Basile et al. 2019), para la detección de discurso de odio contra inmigrantes y mujeres en mensajes de Twitter para idioma español e inglés o HaterNet, un sistema diseñado en colaboración con la Oficina Nacional de Lucha Contra los Delitos de Odio de la Secretaría de Estado de Seguridad de España (perteneciente al Ministerio del Interior), que busca identificar y monitorear la evolución del discurso de odio en redes sociales utilizando el enfoque de una red neuronal LSTM + MLP (Pereira-Kohatsu et al. 2019).

3. Relevamiento de datasets disponibles

Para este estudio, se han explorado diversos datasets disponibles en repositorios como Google Dataset Search, Hugging Face, Kaggle y Zenodo¹⁰, que cubren una amplia variedad de temas y enfoques en la clasificación de lenguaje tóxico en español. Algunas de las palabras clave y filtros utilizados para este relevamiento fueron: *toxic language, hate speech, offensive language, spanish, social bias, nlp, classification tasks*.

En la *Tabla A1*, presente en los [Anexos](#), puede encontrarse el listado completo de datasets relevados, de los cuales destacamos aquí solo algunos:

- OfendES¹¹: Un dataset de comentarios en redes sociales en español que clasifica comentarios ofensivos, utilizado en la competición IberLEF 2021.
- NewsCom-TOX¹²: Un corpus de comentarios en noticias sobre inmigración, utilizado en la competición DETOXIS.
- Spanish Hate Speech Superset¹³: Un dataset que combina varios conjuntos de datos relacionados con hate speech en español.

¹⁰ <https://datasetsearch.research.google.com/>, <https://huggingface.co/datasets>, <https://www.kaggle.com/datasets>, <https://zenodo.org/>

¹¹ <https://huggingface.co/datasets/fmplaza/offendes>

¹² <https://detoxisiberlef.wixsite.com/website/corpus>

¹³ <https://huggingface.co/datasets/manueltonneau/spanish-hate-speech-superset>

Uno de los elementos clave en este trabajo es la generación de un nuevo dataset a partir de la fusión de datasets ya existentes, incluyendo además la anotación de la variedad lingüística regional a la que pertenecen (español latinoamericano o español peninsular).

4. Relevamiento de recursos disponibles y modelos aplicables

Además de los modelos pre-entrenados para el idioma español como BETO¹⁴, BERTIN¹⁵ y RoBERTuito¹⁶, se han identificado otros recursos y modelos disponibles en repositorios de código abierto como Hugging Face, Github o Kaggle¹⁷. Estos incluyen librerías, herramientas y modelos enfocados en la detección de lenguaje ofensivo y de odio en español, como:

- Algoritmo de detección de expresiones de odio en español¹⁸, del proyecto HateMedia.
- EDIA: Estereotipos y Discriminación en Inteligencia Artificial¹⁹
- Hate Speech Library in Spanish²⁰
- Hate-speech-spanish-lexicons²¹

A través de la revisión de la literatura existente, se observa que aunque ha habido un considerable progreso en la detección de la toxicidad en redes sociales, aún persisten desafíos significativos. La mayoría de los enfoques se centran en el idioma inglés, lo que crea una brecha en la investigación sobre la toxicidad en otros idiomas, como el español. Además, muchos modelos no abordan adecuadamente el sesgo cultural y lingüístico, lo que limita su aplicabilidad en contextos multilingües.

Este proyecto busca cerrar esta brecha al aplicar modelos de aprendizaje automático y NLP en un conjunto de datos de tweets en español, considerando el contexto regional y cultural. Esto no solo contribuye al campo del análisis de sentimiento, sino que también establece un precedente para el desarrollo de herramientas más precisas y adaptativas para abordar la toxicidad en el discurso en línea.

¹⁴ <https://github.com/dccuchile/beto>

¹⁵ <https://huggingface.co/bertin-project/bertin-roberta-base-spanish>

¹⁶ <https://github.com/pysentimiento/robertuito>

¹⁷ <https://huggingface.co/>, <https://github.com/>, <https://www.kaggle.com/>

¹⁸ <https://github.com/esaidh266/Algorithm-for-detection-of-hate-speech-in-Spanish/tree/main>

¹⁹ <https://huggingface.co/spaces/vialibre/edia>

²⁰ https://data.niaid.nih.gov/resources?id=zenodo_11099511

²¹ <https://huggingface.co/datasets/SINAI/hate-speech-spanish-lexicons>

OBJETIVOS

Trazando el camino: objetivos del proyecto para combatir la toxicidad online

Este proyecto tiene como finalidad abordar la problemática de la detección de contenidos tóxicos a través del procesamiento de lenguaje natural en español. Los objetivos se dividen en generales y específicos, orientando la investigación hacia la identificación, evaluación y mitigación de sesgos regionales en los modelos utilizados.

Objetivo general

Desarrollar un modelo de clasificación para identificar tweets tóxicos en español, que considere el sesgo lingüístico y regional entre los tweets provenientes de Latinoamérica y España. Este objetivo busca proporcionar una solución efectiva y contextualizada para la detección de la toxicidad en el discurso en línea, contribuyendo a la mejora de herramientas de moderación en plataformas digitales y fomentando un entorno más saludable en las redes sociales.

Objetivos específicos

1. Realizar un análisis exhaustivo de los datasets disponibles: Teniendo en cuenta la relevancia de los datos etiquetados para el entrenamiento de los modelos de clasificación automática, el primer objetivo específico es realizar un relevamiento de los datasets disponibles actualmente para el análisis del lenguaje tóxico en español, analizar sus características y seleccionar aquellos de utilidad para este caso en concreto. La revisión incluye tanto repositorios abiertos como investigaciones académicas previas. El objetivo es seleccionar datasets que abarquen comentarios de diversas regiones hispanohablantes (Latinoamérica y España) para asegurar una representación adecuada de las variantes regionales del idioma.

2. Conseguir un nuevo dataset con formato adecuado y uniforme: Una vez seleccionados los datasets, se buscará fusionarlos en un formato coherente y uniforme con el objetivo de garantizar un dataset no utilizado hasta el momento, que nos pueda aportar información relevante y novedosa para la clasificación e identificación de lenguaje tóxico

3. Implementar y evaluar distintos modelos para la clasificación de tweets tóxicos: El tercer objetivo específico es implementar y comparar varios modelos de clasificación que abarquen enfoques tradicionales de ML y enfoques de DL. Cada uno de estos modelos será entrenado y evaluado con el objetivo de determinar cuáles de estos enfoques son más efectivos para este tipo de tarea, teniendo en cuenta su complejidad y aplicabilidad práctica.

4. Analizar la correlación entre el rendimiento del modelo y la variable de región: Este objetivo específico tiene como finalidad analizar si los modelos presentan resultados diversos en función de la región de origen de los tweets (Latinoamérica vs. España). El objetivo es identificar posibles discrepancias en el rendimiento de los modelos según la región y evaluar si existe una correlación entre la variable región y el rendimiento del modelo en la clasificación de tweets tóxicos. A través de este análisis, se buscará proponer estrategias para mitigar el

sesgo regional en los modelos de detección de toxicidad, contribuyendo a la creación de herramientas más justas e inclusivas.

SOLUCIÓN PLANTEADA

Estrategias de detección: metodología propuesta para la clasificación de tweets ofensivos

METODOLOGÍA

Para abordar el problema de la identificación de tweets tóxicos y el análisis de sesgos relacionados con la región, se ha adoptado una metodología estructurada en varias etapas. Esta metodología sigue un enfoque iterativo y se basa en prácticas validadas en el campo del procesamiento de lenguaje natural y del aprendizaje automático.

Etapas 1: Recolección de datos y análisis inicial

Recopilación de datos mediante el uso de repositorios y bases de datos públicas. Se obtuvo un conjunto de datos de tweets etiquetados, incluyendo detalles sobre sus dimensiones, origen del contenido textual, definición de etiquetas de toxicidad y metadatos como la región.

Etapas 2: Preprocesamiento de datos

Esta etapa se centró en limpiar y preparar los datos para el análisis. Incluyó la eliminación de datos duplicados y la normalización de los textos, así como la tokenización, eliminación de stopwords y la lematización. También se manejaron emojis y otros caracteres especiales. Este paso es fundamental para garantizar que los modelos de clasificación trabajen con datos limpios y representativos del lenguaje.

Etapas 3: Análisis exploratorio de datos (EDA)

Se realizó un análisis exploratorio para entender la distribución de las etiquetas de toxicidad y la influencia de la región en los datos. Se utilizaron gráficos y estadísticas descriptivas para identificar patrones y tendencias relevantes.

Etapas 4: Modelado y evaluación

En esta etapa, se implementaron cuatro modelos diferentes de clasificación (Regresión Logística, Random Forest, SVM y BERT), ajustando los hiper-parámetros para maximizar su rendimiento y ajustar el desbalance de clases. Se aplicaron métricas de evaluación estándar, como precisión, recall y F1-score para identificar el modelo más efectivo para este tipo de análisis.

Etapas 5: Análisis de resultados y conclusiones

Finalmente, se realizó un análisis de los resultados obtenidos para evaluar el rendimiento de los modelos y su relación con la variable de región. Se extrajeron conclusiones y se plantearon posibles líneas de investigación futura.

DESARROLLO DE CADA ETAPA

La solución propuesta para abordar el problema de la clasificación de comentarios tóxicos en español se basa en una metodología integral que combina la exploración de datos, el uso de técnicas avanzadas de procesamiento de lenguaje natural y la aplicación de modelos de machine learning y transformers. Esta sección detalla cada uno de los componentes de la solución planteada, así como las herramientas y enfoques específicos que se utilizaron en el desarrollo del proyecto.

En los siguientes enlaces pueden consultarse los notebooks de Google Colab correspondientes tanto la fase inicial del proyecto donde se genera el dataset unificado: [GP_IEBS_Data.ipynb](#) como a las etapas posteriores de preprocesamiento de texto, análisis exploratorio de datos, modelado y evaluación de rendimiento: [GP_IEBS_Modelos_LenguajeTox.ipynb](#)

Etapas 1: Recolección de datos y análisis inicial

La recolección de datos se realizó utilizando repositorios y conjuntos de datos públicos que incluían tweets etiquetados por su contenido tóxico. Se seleccionaron tweets en español, con especial atención a aquellos provenientes de España y Latinoamérica. Esta etapa fue crucial para asegurar la diversidad y representatividad del dataset.

De los 21 datasets analizados (ver detalles en la Tabla A1 disponible en [Anexos](#)) se decidió trabajar con 6 debido a sus características uniformes (contenido textual de tweets, etiquetados para identificación de lenguaje tóxico, con metadatos disponibles acerca de la región de proveniencia).

Los datasets seleccionados fueron:

- **HaSCoSva**: 4000 tweets etiquetados correspondientes las dos variantes del idioma español
- **Multilingual Hate Speech**: 4831 tweets etiquetados correspondientes a diversos países
- **MEX_A3T**: 11000 tweets etiquetados para lenguaje ofensivo provenientes de México
- **Chileno**: 9834 tweets etiquetados para lenguaje ofensivo provenientes de Chile
- **StereoHoax-ES**: 5349 tweets etiquetados para lenguaje ofensivo provenientes de España
- **Gender Bias in Spanish Tweets**: 1914 tweets etiquetados para lenguaje ofensivo provenientes de España

Posteriormente, se cargaron los diferentes archivos de cada dataset, se uniformizó su estructura y se fusionaron en un único archivo csv conteniendo los siguientes campos:

- **text**: contenido del tweet, en formato string
- **region**: región de donde proviene el tweet, en formato binario: 0 (para Latinoamérica) y 1 (para España)
- **dtst**: dataset de origen, en formato string
- **label**: etiqueta de clasificación, en formato binario: 0 (para contenido no tóxico) y 1 (para contenido tóxico)

Se procedió entonces a cargar el archivo y leerlo en un dataframe de pandas para poder así realizar un análisis exploratorio inicial de los datos (EDA sin limpieza de textos).

Tamaño y distribución de las etiquetas y regiones: En este paso, se visualizó el tamaño general del dataset unificado (correspondiente a 29761 registros y 4 columnas), y la distribución de las etiquetas *label* y la variable *region*. Esto permitió identificar que el dataset está desbalanceado, ya que hay más tweets clasificados como no ofensivos (63%) que tweets ofensivos (37%), tal como se muestra en la *Figura 4*. Esta observación es clave, ya que el desbalance puede afectar el rendimiento de los modelos de clasificación. Asimismo, se comprobó la distribución de los tweets por región para tener un panorama de la procedencia geográfica de los datos.

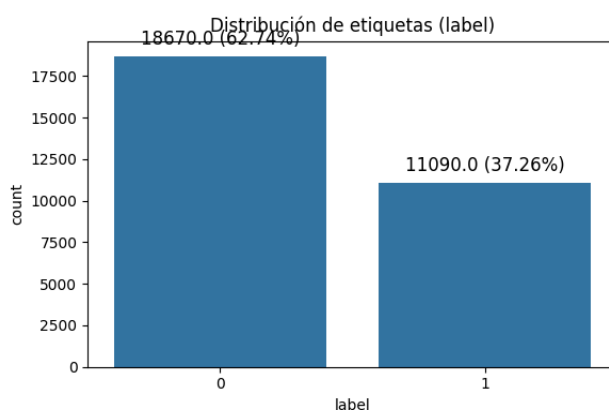


Figura 4. Distribución de etiquetas. Fuente propia

Al analizar la relación entre las etiquetas y regiones (*Figura 5*), observamos también una menor distribución de tweets calificados como tóxicos para la región de España, lo que en definitiva puede afectar a los resultados de los análisis por región:

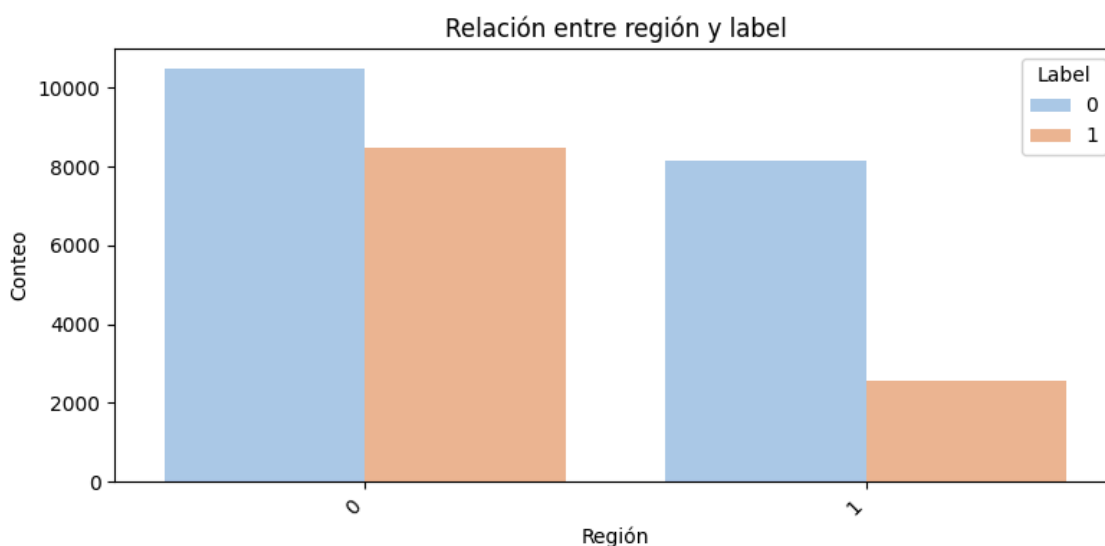


Figura 5. Distribución de etiquetas por región. Fuente propia

Longitud de los tweets: Se calculó la longitud de cada tweet para analizar su distribución. A través de este análisis inicial, se descubrió que existen outliers, es decir, algunos tweets excepcionalmente largos, que podrían distorsionar el análisis y los modelos. Por lo tanto, fue importante observar cómo estos tweets impactan la distribución general (*Figura 6*).

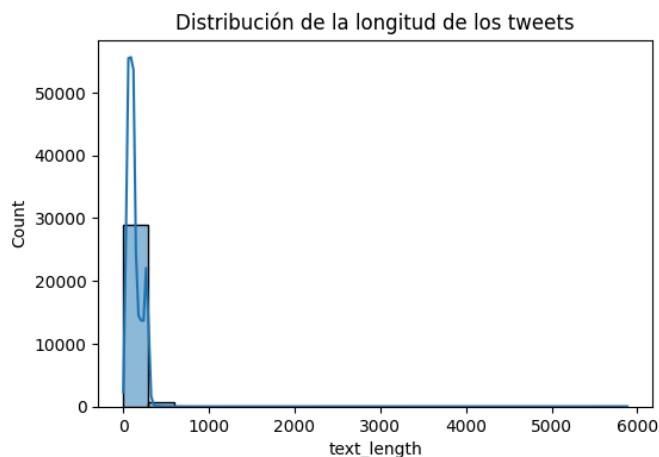


Figura 6. Longitud de tweets. Fuente propia

Resumen estadístico: Se generó un resumen estadístico para obtener una visión cuantitativa de las características del dataset, como la longitud mínima, media y máxima de los tweets, así como la dispersión en la longitud de los mismos.

Longitud de los tweets sin outliers y limpieza de outliers: Dado que se detectaron tweets anormalmente largos (outliers), se decidió eliminarlos utilizando el rango intercuartílico (IQR). Esta técnica ayuda a identificar y eliminar valores que están fuera de los límites esperados. Después de la eliminación de los outliers, la longitud de los tweets se ajustó a una distribución más razonable para el análisis (*Figura 6*).

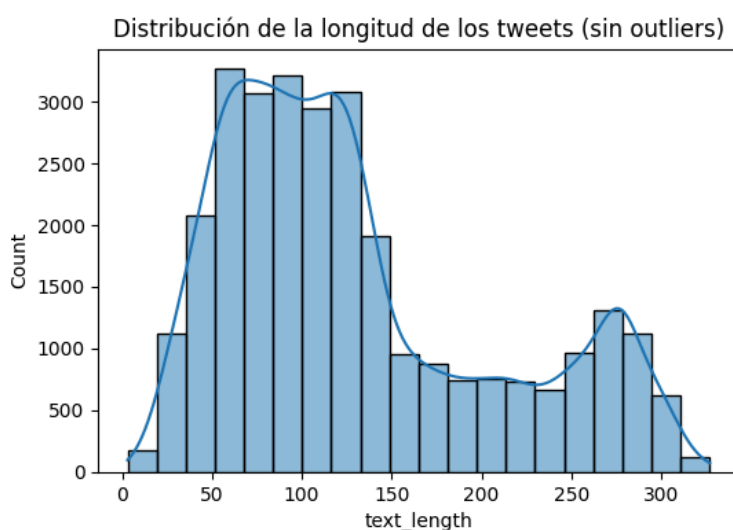


Figura 6. Longitud de tweets (sin outliers). Fuente propia

Resumen estadístico (sin outliers): Tras eliminar los outliers, se generó un nuevo resumen estadístico para verificar cómo la eliminación de estos valores extremos afectó las medidas estadísticas. Esto permitió obtener una visión más precisa de las características de los tweets "normales" en cuanto a longitud.

Etapas 2: Preprocesamiento de datos

El preprocesamiento de los tweets se realizó utilizando la biblioteca spaCy²², ya que cuenta con características específicas para el idioma español (modelo: es_core_news_sm). En este paso, los tweets fueron limpiados y transformados de la siguiente manera:

- **Normalización del texto:** Para uniformizar los datos, se convirtieron todos los textos a minúsculas y se eliminaron urls, nombres de usuarios y caracteres no alfabéticos, que no aportan información relevante al análisis de texto.
- **Manejo de emojis y símbolos:** Los emojis fueron convertidos a su descripción textual, permitiendo que su significado fuera incluido en el análisis.
- **Tokenización y stopwords:** Los tweets fueron divididos en tokens (palabras individuales), eliminando las stopwords (palabras vacías comunes como preposiciones y artículos que no aportan valor semántico, como por ejemplo "el", "la", "a", etc.), que no son útiles para el análisis.
- **Stemming/Lematización:** Se realizó la lematización, es decir, se convirtió cada palabra a su forma base, lo que ayuda a mejorar el rendimiento en tareas de análisis de texto al tratar palabras con el mismo significado como equivalentes. Se utilizó el lematizador de spaCy para asegurar una correcta reducción.
- **Limpieza de datos duplicados y NaN:** Finalmente, se eliminaron los registros duplicados y se trató cualquier dato faltante.

Este preprocesamiento fue crucial para asegurar que los tweets estuvieran en un formato adecuado para su análisis y modelado posterior.

Etapas 3: Análisis exploratorio de datos (EDA)

Durante la exploración de los datos ya procesados, se emplearon herramientas como Pandas y Matplotlib para visualizar la distribución de longitud de los tweets, palabras más frecuentes y nubes de palabras:

Distribución de la longitud de los tweets después del preprocesamiento: Una vez que los tweets fueron preprocesados, se volvió a calcular y visualizar la distribución de la longitud de los mismos. Este paso permitió observar cómo el preprocesamiento afectó la longitud de los

²² <https://spacy.io/models/es>

tweets, ya que después de eliminar elementos como URLs y caracteres especiales, los tweets se acortaron considerablemente.

Resumen estadístico: Se generó un nuevo resumen estadístico de los tweets procesados, proporcionando información detallada sobre la longitud de los textos después del preprocesamiento. Esto ayudó a verificar que el proceso de limpieza y lematización había reducido de forma efectiva la longitud de los tweets.

Palabras más frecuentes: Se realizó un análisis de frecuencia de palabras para identificar cuáles eran las palabras más comunes en los tweets procesados. Esto permitió observar patrones en el lenguaje utilizado, identificando qué términos aparecían con mayor regularidad en el conjunto de datos.

Nube de palabras: Además del análisis de frecuencia, se generó una nube de palabras para visualizar gráficamente las palabras más comunes de manera atractiva y fácil de interpretar. Las palabras más frecuentes se muestran en tamaños más grandes, lo que facilita una rápida identificación de los términos más relevantes.

Este mismo análisis se replicó para cada región:

Distribución de la longitud de los tweets por región: Se analizó la distribución de la longitud de los tweets para cada región de forma separada. Este análisis permitió comparar cómo varía la longitud de los tweets dependiendo de la región de origen, proporcionando una visión más granular de las diferencias entre regiones.

Resumen estadístico por región: Se generaron resúmenes estadísticos para cada región, lo que permitió observar cómo las características de los tweets (como la longitud) variaban entre diferentes ubicaciones geográficas. Este análisis detallado por región es importante para detectar posibles sesgos o diferencias en el contenido dependiendo del lugar.

Palabras más frecuentes por región: Se realizó un análisis de las palabras más comunes para cada región por separado. Esto permitió identificar qué términos son más utilizados en cada ubicación geográfica, ofreciendo una visión más detallada sobre el lenguaje utilizado en cada región.

Nube de palabras por región: Finalmente, se generaron nubes de palabras para cada región. Esta visualización ayudó a identificar rápidamente qué palabras predominaban en los tweets de cada área geográfica, permitiendo comparar el contenido entre regiones de manera intuitiva.

Etapas 4: Modelado y evaluación

Los modelos para esta tarea se seleccionaron por motivos de rendimiento y accesibilidad siendo los elegidos tres modelos de ML y uno de DL. Se entrenaron luego utilizando el conjunto de datos preprocesado. Cada modelo fue evaluado mediante validación cruzada y se utilizaron métricas de precisión, recall y F1-score para medir su rendimiento.

Los modelos seleccionados finalmente fueron:

- **Regresión Logística (LR):** Un modelo simple pero efectivo que sirve como base para comparaciones.
- **Random Forest (RF):** Se utilizó este modelo para capturar interacciones no lineales en los datos.
- **Support Vector Machine (SVM):** Este modelo se eligió por su eficacia en problemas de clasificación con un margen claro de separación.
- **DistilBERT²³:** Se implementó este modelo de aprendizaje profundo para abordar la complejidad del lenguaje y mejorar la detección de contexto en los tweets. En un principio se intentó aplicar modelos de BERT más grandes, como RoBERTa o BERT, o realizar fine-tuning adicional sobre el modelo, pero por motivos de falta de recursos computacionales se decidió implementar solo este modelo más ligero y accesible.

4.1 Importación de librerías y preparación de datos

En esta etapa inicial, se importaron todas las librerías necesarias para la creación, ajuste y la evaluación del rendimiento de los modelos.

El siguiente paso fue la preparación de los datos para su uso en los modelos. Esta etapa incluyó las siguientes acciones clave:

División del dataset:

El dataset de tweets fue dividido en dos partes: “X” (tweets preprocesados) como características de entrada y “y” (*label*, o clase binaria) como la variable objetivo que se quiere predecir.

Se utilizó la función `train_test_split` de sklearn²⁴ para dividir los datos en un conjunto de entrenamiento y prueba. El conjunto de entrenamiento se utilizó para ajustar los modelos, mientras que el conjunto de prueba se mantuvo separado para la evaluación del rendimiento del modelo. El conjunto de prueba fue el 20% del total de los datos.

Representación de los textos:

Los tweets preprocesados fueron transformados en representaciones numéricas utilizando TF-IDF (Term Frequency-Inverse Document Frequency). Esta técnica convierte el texto en vectores numéricos que pueden ser utilizados por los modelos de machine learning, considerando no solo la frecuencia de las palabras en el documento, sino también su frecuencia en todo el corpus. Se limitaron a 5000 características para reducir la dimensionalidad.

Esta preparación de los datos es crucial para poder aplicar modelos de ML, ya que permite transformar texto en una representación numérica utilizable.

²³ https://huggingface.co/docs/transformers/en/model_doc/distilbert

²⁴ <https://scikit-learn.org/stable/>

Función para evaluar los modelos:

Para evaluar cada modelo, definimos una función llamada *evaluate_model*. Esta función entrena el modelo en el conjunto de entrenamiento y luego lo evalúa en el conjunto de prueba utilizando varias métricas de rendimiento, como la matriz de confusión, el reporte de clasificación (que incluye precisión, recall y f1-score) y la AUC-ROC.

4.2 Modelos estadísticos de Machine Learning

Se inicia con modelos clásicos de ML para establecer una línea base de rendimiento en la clasificación de sentimientos. Se aplicaron tres modelos diferentes para predecir la clase binaria: Logistic Regression, SVM, y Random Forest.

Los tres modelos se entrenaron manejando el desbalance de clases mediante la opción *class_weight='balanced'*, que ajusta automáticamente los pesos de las clases de forma inversamente proporcional a su frecuencia.

Cada uno de los tres modelos fue entrenado con el conjunto de datos tal como estaba, es decir, sin generar instancias sintéticas de la clase minoritaria (tweets tóxicos). El desbalance fue manejado solo ajustando los pesos internos de los modelos. Este enfoque tiene la ventaja de mantener el tamaño original del dataset, sin modificar el conjunto de datos, pero a veces no es suficiente para manejar adecuadamente un desbalance de clases severo.

Después de entrenar los modelos, se realizaron predicciones sobre el conjunto de prueba para evaluar su rendimiento.

4.2.1 Regresión Logística

La regresión logística utiliza características predictoras para construir un modelo lineal que estima la probabilidad de que una observación pertenezca o no a una clase determinada. Se crea así un modelo lineal que estima las probabilidades de una instancia perteneciente a una clase e intenta separar las clases tóxicas/no tóxicas mediante una frontera lineal.

4.2.2 Support Vector Machine (SVM)

Un modelo que intenta encontrar un hiperplano que separe de manera óptima las dos clases. Las SVM son clasificadores cuyo resultado se basa en una frontera de decisión generada por vectores de soporte, es decir, por los puntos más cercanos a la frontera de decisión. La forma de la frontera está determinada por una función kernel. De esta manera, es posible resolver problemas que no se pueden resolver mediante una frontera lineal. Intuitivamente, una buena separación se logra mediante el hiperplano que tiene la mayor distancia a los vectores de soporte, ya que, en general, cuanto mayor es el margen menor es el error de generalización del clasificador.

4.2.3 Random Forest

Un modelo basado en árboles de decisión que crea múltiples árboles y promedia sus resultados para mejorar la predicción. Estos bosques aleatorios (RF) funcionan entonces como

una serie de clasificadores de conjunto cuya construcción se basa en el uso de varios árboles de decisión donde cada árbol se entrena con una muestra extraída del conjunto de datos de entrenamiento y utilizando un subconjunto aleatorio de las características. Los árboles de decisión dividen el espacio factorial según pruebas de valor, lo que da como resultado una clasificación no lineal. Los nodos de los árboles se determinan de manera que se maximice la ganancia de información. Existen diferentes criterios para determinar esta ganancia de información, siendo Gini y la entropía dos de los más comunes.

4.3 Modelos de Deep Learning con transformers

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de procesamiento de lenguaje natural basado en la arquitectura de transformers, diseñado para comprender el contexto de una palabra en función de las palabras que la rodean, tanto antes como después. Los transformers son una arquitectura de red neuronal que utiliza un mecanismo de atención para procesar secuencias de datos, como texto, de manera eficiente y paralela, capturando relaciones contextuales entre palabras a largo plazo (Vaswani et al. 2023). BERT utiliza embeddings contextuales, que son representaciones numéricas de palabras que varían según el contexto en el que aparecen, lo que permite generar interpretaciones más precisas del significado de las palabras en diferentes oraciones y escenarios (Devlin et al. 2019).

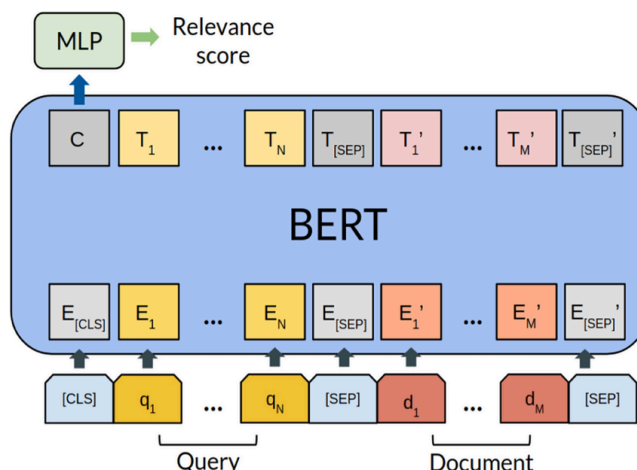


Figura 7. Arquitectura BERT. Tomado de Devlin et al. (2019)

4.3.1 DistilBERT

Para el modelo de DistilBERT, un modelo de Transformers que ha sido pre-entrenado en grandes cantidades de datos textuales, seguimos una metodología diferente a la empleada con los modelos de ML. DistilBERT ya está entrenado en el conocimiento del lenguaje español, por lo que lo utilizamos para la clasificación de los tweets tóxicos y no tóxicos.

En este paso, utilizamos el tokenizador de DistilBERT para convertir los tweets en los tokens que el modelo puede entender. También calculamos los pesos de clase para manejar el desbalance de clases y personalizamos el Trainer para que utilice estos pesos durante el entrenamiento.

4.4 Evaluación de modelos

La evaluación de los modelos se lleva a cabo mediante el cálculo de las siguientes métricas de rendimiento para evaluar la efectividad de los modelos en la clasificación de sentimientos en general: precisión, recall, F1-score, accuracy, matriz de confusión, curva ROC y área bajo la curva AUC (ver sección de [Evaluación](#)).

Etapas 5: Análisis de resultados y conclusiones

Se analizaron los resultados obtenidos en términos de rendimiento de los modelos y su relación con la variable de región. Se identificaron diferencias en la efectividad de los modelos según el origen geográfico de los tweets, sugiriendo la necesidad de personalizar los modelos para diferentes contextos (Ver secciones de [Resultados](#) y [Conclusiones](#)).

La solución planteada se fundamenta en una metodología estructurada que permite abordar el problema de la toxicidad en el discurso en línea de manera efectiva. Al combinar técnicas de procesamiento de lenguaje natural con modelos de aprendizaje automático y profundo, este proyecto busca contribuir a la detección de contenido ofensivo, sino que también abre nuevas líneas de investigación en la comprensión del sesgo regional en el lenguaje.

EVALUACIÓN

Midiendo los modelos: evaluación y métricas de rendimiento

La evaluación de los modelos desarrollados es un componente crucial para determinar la eficacia en la clasificación de textos y debe seguir una serie de metodologías estructuradas y validadas, asegurando que los resultados obtenidos sean confiables, replicables y útiles. A continuación, se detalla el enfoque seguido para la evaluación de este estudio.

1. Metodología de evaluación

La evaluación de los modelos de clasificación se realizó en varias etapas clave, cada una de las cuales aborda aspectos diferentes de la solución:

1.1. Definición de métricas de evaluación

Para evaluar el rendimiento de los modelos, se definieron las siguientes métricas clave:

Precisión (Precision): La precisión mide la proporción de verdaderos positivos (tweets tóxicos correctamente identificados) frente a todos los tweets que fueron predichos como tóxicos (verdaderos positivos + falsos positivos). En términos simples, responde a la pregunta: *¿De todos los tweets clasificados como tóxicos, cuántos realmente lo son?*.

Fórmula:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Positivos (FP)}}$$

Recall (Sensibilidad o Tasa de Detección): El recall mide la proporción de verdaderos positivos frente a todos los tweets que en realidad son tóxicos (verdaderos positivos + falsos negativos). Responde a la pregunta: *¿De todos los tweets que son realmente tóxicos, cuántos fueron correctamente identificados por el modelo?*.

Fórmula:

$$\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

F1-Score: El f1-score es la media armónica entre la precisión y el recall, y proporciona un balance entre ambas métricas. Es especialmente útil cuando existe un desbalance de clases, como en nuestro caso (más tweets no tóxicos que tóxicos). Un f1-score más alto indica un mejor balance entre la precisión y el recall.

Fórmula:

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Accuracy (Exactitud): La exactitud mide el porcentaje de predicciones correctas sobre el total de predicciones. Indica qué tan bien el modelo clasifica correctamente tanto los tweets tóxicos como los no tóxicos.

Fórmula:

$$\text{Exactitud} = \frac{\text{Predicciones Correctas}}{\text{Total de Predicciones}}$$

AUC-ROC (Área bajo la curva ROC): El AUC-ROC mide la capacidad de un modelo para distinguir entre las clases (tóxicos y no tóxicos). ROC es una curva que muestra el trade-off entre el True Positive Rate (Sensibilidad) y el False Positive Rate. El AUC (Área bajo la curva) es un valor entre 0 y 1, donde un valor más cercano a 1 indica que el modelo es mejor en la distinción entre clases. Es decir que si se da que:

- ❖ AUC = 1: El modelo clasifica perfectamente.
- ❖ AUC = 0.5: El modelo clasifica aleatoriamente.
- ❖ AUC < 0.5: El modelo clasifica peor que aleatoriamente.

Matriz de Confusión: La matriz de confusión es una tabla que permite visualizar el rendimiento de un modelo de clasificación. Muestra las verdaderas clases versus las predicciones hechas por el modelo. Contiene:

- ❖ Verdaderos Positivos (TP): Tweets tóxicos correctamente clasificados.
- ❖ Verdaderos Negativos (TN): Tweets no tóxicos correctamente clasificados.
- ❖ Falsos Positivos (FP): Tweets no tóxicos incorrectamente clasificados como tóxicos.
- ❖ Falsos Negativos (FN): Tweets tóxicos incorrectamente clasificados como no tóxicos.

La matriz de confusión ayuda a evaluar no solo la precisión general del modelo, sino también sus errores en términos de falsos positivos y falsos negativos, que pueden tener implicaciones importantes en la moderación de contenido online.

1.2. División en datos de entrenamiento y test

El conjunto de datos se dividió en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%). El conjunto de entrenamiento se utilizó para ajustar los modelos y el conjunto de prueba se utilizó para evaluar su rendimiento. Esta separación es crucial para garantizar que el rendimiento medido no se vea influenciado por datos que el modelo ya ha visto.

2. Implementación de los modelos

Se implementaron y evaluaron los cuatro modelos de clasificación:

- **Regresión Logística**
- **Random Forest**
- **Support Vector Machines (SVM)**
- **DistilBERT**

Cada modelo se ajustó y evaluó utilizando las métricas definidas anteriormente, proporcionando una visión clara de su rendimiento.

3. Análisis de resultados

Los resultados se analizaron para identificar patrones y tendencias. Se evaluó si los modelos tenían un rendimiento desigual entre los tweets provenientes de Latinoamérica y Europa.

3.1. Comparación de modelos

Una vez evaluados los diferentes modelos (incluyendo aquellos basados en ML tradicional y los basados en transformers), se lleva a cabo una comparación sistemática de su rendimiento, resaltando aquellos que mostraron mejores resultados en términos de precisión, recall y F1-Score. Se incluyeron análisis de la matriz de confusión para identificar áreas de mejora en cada modelo.

3.2. Evaluación por región

Se lleva a cabo un análisis específico sobre cómo el campo *'region'* puede afectar el rendimiento de los modelos. Se evaluó si los modelos tendían a clasificar incorrectamente más tweets de una región que de otra y se propusieron soluciones para mitigar cualquier sesgo identificado.

4. Resultados y reflexiones finales

En la siguiente sección se presentan los resultados obtenidos con sus implicaciones en el contexto de la detección y moderación del discurso tóxico.

Finalmente, se incluyen recomendaciones basadas en los resultados de la evaluación. Estas recomendaciones están dirigidas a futuras implementaciones de modelos de clasificación de tweets, enfatizando la importancia de la consideración del sesgo regional y lingüístico.

RESULTADOS

Revelaciones del análisis: resultados clave y observaciones de la clasificación de tweets

En esta sección, se describen los resultados obtenidos a partir de la evaluación de la solución planteada, centrándose en la clasificación de tweets tóxicos. Se presentarán los resultados en función de las métricas definidas anteriormente, así como un análisis del rendimiento de cada modelo y una evaluación específica en relación con la región de origen de los tweets.

1. Desempeño de los modelos

A continuación se describen los resultados obtenidos para cada modelo de clasificación en el dataset de tweets, tanto de forma global como desglosados por región (Latinoamérica y España). Estos resultados permiten evaluar la capacidad de cada modelo para clasificar correctamente tweets como tóxicos o no tóxicos, y analizar si presentan diferencias en su rendimiento dependiendo de la región.

1.1. Logistic Regression

Matriz de confusión:

- **Clase 0 (no tóxico):** 3030 predicciones correctas y 693 incorrectas.
- **Clase 1 (tóxico):** 1631 predicciones correctas y 579 incorrectas.

Reporte de clasificación:

- **Precisión global (Accuracy):** 0.79
- **Precision:** 0.84 (para clase 0, no tóxicos), 0.70 (para clase 1, tóxicos)
- **Recall:** 0.81 (clase 0), 0.74 (clase 1)
- **F1-score:** 0.83 (clase 0), 0.72 (clase 1)
- **AUC-ROC:** 0.85

La regresión logística muestra un desempeño sólido con una precisión general del 79% y un AUC-ROC de 0.85, lo que sugiere que el modelo distingue bien entre tweets tóxicos y no tóxicos. Sin embargo, el f1-score para la clase tóxica es más bajo (0.72), lo que indica que el modelo tiene más dificultad en identificar correctamente los tweets tóxicos en comparación con los no tóxicos.

Resultados por región:

Latinoamérica: Accuracy: 0.81, AUC-ROC: 0.89

- El modelo muestra un buen rendimiento en Latinoamérica, con un f1-score de 0.80 para la clase tóxica, lo que indica una capacidad razonable para clasificar tweets tóxicos.

España: Accuracy: 0.84, AUC-ROC: 0.88

- En España, la regresión logística también tiene un buen rendimiento, pero el f1-score para la clase tóxica es más bajo (0.66), lo que indica que el modelo tiene más dificultades para clasificar correctamente los tweets tóxicos en esta región.

El modelo funciona de manera similar en ambas regiones, pero el rendimiento es ligeramente mejor en Latinoamérica, especialmente en la clasificación de tweets tóxicos.

1.2. Support Vector Machine (SVM)

Matriz de confusión:

- **Clase 0 (no tóxico):** 3153 predicciones correctas y 570 incorrectas.
- **Clase 1 (tóxico):** 1552 predicciones correctas y 658 incorrectas.

Reporte de clasificación:

- **Precisión global (Accuracy):** 0.79
- **Precision:** 0.83 (clase 0), 0.73 (clase 1)
- **Recall:** 0.85 (clase 0), 0.70 (clase 1)
- **F1-score:** 0.84 (clase 0), 0.72 (clase 1)
- **AUC-ROC:** 0.86

El modelo SVM presenta una precisión global similar a la de la regresión logística (79%) y un AUC-ROC ligeramente superior (0.86). Esto sugiere que SVM es mejor que la regresión logística para distinguir entre clases, aunque sigue teniendo problemas para clasificar los tweets tóxicos correctamente.

Resultados por región:

Latinoamérica: Accuracy: 0.91, AUC-ROC: 0.97

- El modelo SVM es significativamente más efectivo en Latinoamérica, con una precisión del 91% y un f1-score de 0.91 para la clase tóxica, lo que indica un excelente rendimiento.

España: Accuracy: 0.94, AUC-ROC: 0.97

- En España, el SVM también obtiene un rendimiento muy alto, con una precisión del 94% y un f1-score de 0.86 para la clase tóxica.

El modelo SVM sobresale en ambas regiones, con una ligera ventaja en España debido a una mejor precisión en la clasificación de tweets no tóxicos. En general, es uno de los modelos más equilibrados en cuanto a precisión y recall para ambas clases.

1.3. Random Forest

Matriz de confusión:

- **Clase 0 (no tóxico):** 3208 predicciones correctas y 515 incorrectas.
- **Clase 1 (tóxico):** 1443 predicciones correctas y 767 incorrectas.

Reporte de clasificación:

- **Precisión global (Accuracy):** 0.78
- **Precision:** 0.81 (clase 0), 0.74 (clase 1)
- **Recall:** 0.86 (clase 0), 0.65 (clase 1)
- **F1-score:** 0.83 (clase 0), 0.69 (clase 1)
- **AUC-ROC:** 0.84

El modelo Random Forest obtiene una precisión global del 78% y un AUC-ROC de 0.84, siendo más eficiente en clasificar correctamente los tweets no tóxicos. El recall para la clase tóxica (0.65) es el más bajo entre los modelos evaluados, lo que indica que no es tan bueno identificando todos los tweets tóxicos.

Resultados por región:

Latinoamérica: Accuracy: 0.95, AUC-ROC: 0.98

- Random Forest sobresale en Latinoamérica con una precisión del 95% y un f1-score de 0.94 para la clase tóxica, lo que muestra un rendimiento excepcional.

España: Accuracy: 0.96, AUC-ROC: 0.98

- En España, el modelo también es muy preciso (96%) con un f1-score de 0.92 para la clase tóxica.

Random Forest muestra un rendimiento notable en ambas regiones, especialmente en la clasificación de tweets tóxicos. La precisión y recall para la clase tóxica son altos en ambos casos, lo que hace de este modelo uno de los más efectivos.

1.4. DistilBERT

Resultados de entrenamiento:

- La pérdida de entrenamiento mostró una disminución constante a lo largo de las épocas, pasando de 0.66 a 0.32.

Matriz de confusión:

- **Clase 0 (no tóxico):** 3114 predicciones correctas y 613 incorrectas.
- **Clase 1 (tóxico):** 1689 predicciones correctas y 521 incorrectas.

Reporte de clasificación:

- **Precisión global (Accuracy):** 0.81
- **Precision:** 0.86 (clase 0), 0.73 (clase 1)
- **Recall:** 0.83 (clase 0), 0.78 (clase 1)
- **F1-score:** 0.85 (clase 0), 0.75 (clase 1)

DistilBERT, nuestro modelo de deep learning, obtiene una precisión global del 81%, con un f1-score de 0.75 para la clase tóxica, lo que indica que es capaz de balancear precisión y recall de manera efectiva.

Resultados por región:

Latinoamérica: Accuracy: 0.79

- DistilBERT en Latinoamérica presenta una precisión del 79% y un f1-score de 0.76 para la clase tóxica, lo que muestra un rendimiento sólido, aunque no tan destacado como otros modelos.

España: Accuracy: 0.83

- En España, el rendimiento de DistilBERT es mayor, con una precisión del 83%, pero tiene dificultades para clasificar correctamente los tweets tóxicos, con un f1-score de 0.57.

DistilBERT funciona mejor en España en términos de precisión, pero su rendimiento en la clasificación de tweets tóxicos en España es inferior al de otros modelos, especialmente SVM y Random Forest.

2. Conclusiones generales

Mejor modelo global:

- Los resultados muestran que el Support Vector Machine (SVM) y DistilBERT son los modelos más prometedores, con una precisión y recall equilibrados en ambas clases. Sin embargo, el modelo DistilBERT sobresale un poco en la detección de la clase tóxica, lo que lo convierte en la mejor opción cuando se busca identificar correctamente el lenguaje ofensivo.
- Logistic Regression y Random Forest también ofrecen un rendimiento aceptable, pero ambos presentan dificultades en la clasificación de la clase tóxica, lo que sugiere que los modelos basados en árboles y en regresión pueden necesitar ajustes adicionales para mejorar en datasets desbalanceados.

Mejor modelo por región:

- *Latinoamérica:* Los modelos tienden a funcionar mejor en esta región, especialmente en la clase no tóxica. SVM y Random Forest destacan por su alta precisión y recall en

ambas clases, con valores de AUC-ROC cercanos a 0.98, lo que indica una capacidad sobresaliente para discriminar entre clases.

- *España*: Aunque los resultados son similares, se observa una leve caída en el rendimiento de la clase tóxica, especialmente en DistilBERT, que muestra una disminución en la capacidad de detectar tweets tóxicos. Esto sugiere que las diferencias lingüísticas y culturales entre ambas regiones pueden influir en el rendimiento del modelo.

3. Propuestas de mejora

Ajuste de hiperparámetros:

- Logistic Regression: Ajustar parámetros como la regularización (parámetro C) y explorar el uso de otras funciones de regularización, como L1 o L2. Esto podría mejorar la generalización del modelo y su capacidad para manejar el desbalance de clases.
- SVM: Mejorar el ajuste de parámetros como C (regularización) y gamma (coeficiente de kernel) para optimizar el margen entre las clases. Usar técnicas de búsqueda en cuadrícula (grid search) o búsqueda aleatoria (random search) para encontrar los mejores valores.
- Random Forest: Ajustar la cantidad de árboles en el bosque (n_estimators), la profundidad máxima del árbol (max_depth) y el número mínimo de muestras por división (min_samples_split). También se podría usar bagging o boosting para mejorar el rendimiento en la clase minoritaria.
- DistilBERT: Probar con un mayor número de épocas (num_train_epochs) para un mejor ajuste. También es recomendable ajustar el learning rate para optimizar el proceso de aprendizaje.

Mejora del balance de clases:

- Aunque se utilizaron técnicas como el ajuste de pesos de clase, aplicar técnicas más avanzadas de resampling podría ayudar a equilibrar la representación de las clases tóxicas y no tóxicas. Esto sería especialmente útil en modelos como Logistic Regression y Random Forest, que mostraron dificultades para capturar correctamente la clase minoritaria.
- Las técnicas de sobremuestreo como SMOTE (Synthetic Minority Oversampling Technique) son un ejemplo ya que servirían para aumentar el número de ejemplos de la clase minoritaria (tweets tóxicos) generando nuevos ejemplos sintéticos, lo que podría ayudar a mejorar la precisión en la clase minoritaria y evitar que el modelo ignore estos ejemplos.

Uso de embeddings más robustos:

- Aunque TF-IDF es una técnica efectiva en modelos tradicionales, explorar el uso de embeddings contextuales como Word2Vec o FastText podría mejorar la representación de los textos, especialmente para capturar el significado en diferentes dialectos y variantes del español.

Modelo ensemble:

- Probar con técnicas de ensemble (combinación de varios modelos) para mejorar el rendimiento general. Un ensamble de Logistic Regression, Random Forest y SVM podría capturar mejor las características de los tweets y mejorar el rendimiento en ambas clases.

Tokenización especializada:

- Usar tokenizadores especializados para el lenguaje español, como el de BETO, en lugar de tokenizadores generales como el de DistilBERT en inglés, para mejorar el rendimiento en tweets en español.

Fine-tuning en modelos de Deep Learning:

- Para modelos como DistilBERT, aumentar el número de épocas y ajustar el learning rate en función del dataset específico. Realizar un fine-tuning más agresivo sobre datos en español para mejorar el rendimiento en tweets de distintas regiones.
- Ajustar los pesos de clase en la función de pérdida para equilibrar el impacto de las clases desbalanceadas.
- Una de las mejoras más inmediatas sería realizar ajustes de los modelos por región. Dado que las diferencias lingüísticas y culturales entre Latinoamérica y España afectan el rendimiento de los modelos, realizar un fine-tuning específico para cada región permitiría adaptar mejor los modelos a las particularidades del lenguaje utilizado en cada contexto. Esto es especialmente relevante para modelos basados en transformers como DistilBERT, que pueden beneficiarse de entrenamientos adicionales en corpus específicos de cada región.

Incorporación de datos adicionales:

- Para mejorar el rendimiento general, una mejora clave sería incorporar datasets adicionales que cubran una mayor variedad de contenido en español, tanto de Latinoamérica como de España. Esto permitiría a los modelos aprender representaciones más generalizables y mejorar su capacidad para manejar variaciones regionales del idioma.
- Implementar técnicas de data augmentation en tweets, como la traducción automática o la paráfrasis de los textos. Estas técnicas generan nuevas versiones de los tweets y pueden ayudar a los modelos a captar mejor la diversidad en el lenguaje.

CONCLUSIONES Y TRABAJOS FUTUROS

Reflexiones finales: conclusiones y rutas para investigaciones futuras

Los resultados globales de los cuatro modelos implementados muestran que DistilBERT y SVM son los más efectivos para la detección de lenguaje tóxico, alcanzando una precisión y recall equilibrados en ambas clases. DistilBERT sobresale en la clase tóxica con un f1-score de 0.75, lo que indica su capacidad superior para identificar lenguaje ofensivo en comparación con los modelos tradicionales como Logistic Regression y Random Forest, que mostraron dificultades para detectar correctamente la clase tóxica. Globalmente, los modelos lograron una precisión de alrededor de 0.79 a 0.81, con AUC-ROC de hasta 0.85.

En cuanto a los resultados por región, los modelos tienden a funcionar mejor en Latinoamérica que en España, especialmente en la detección de tweets no tóxicos. Modelos como SVM y Random Forest alcanzaron un AUC-ROC cercano a 0.98 en Latinoamérica, lo que indica una excelente capacidad de discriminación entre clases. Sin embargo, en España, se observó una ligera disminución en la capacidad de los modelos para detectar tweets tóxicos, particularmente en DistilBERT, donde el recall de la clase tóxica fue más bajo. Estas diferencias sugieren que los factores lingüísticos y culturales regionales impactan en el rendimiento de los modelos, destacando la necesidad de realizar ajustes específicos por región.

1. Relación con los objetivos

Los objetivos establecidos al inicio del proyecto fueron los siguientes:

1. **Objetivo general:** Desarrollar un clasificador efectivo para la identificación de tweets tóxicos, considerando el impacto del idioma y la región de origen en el rendimiento del modelo.
2. **Objetivos específicos:**
 - Realizar un análisis exhaustivo de los datasets disponibles.
 - Preprocesar los datos adecuadamente.
 - Implementar y evaluar distintos modelos de aprendizaje automático y profundo para la clasificación de tweets.
 - Analizar la correlación entre el rendimiento del modelo y la variable de región.

Los resultados obtenidos validan el cumplimiento de estos objetivos, especialmente en la creación de un modelo que no solo clasifica eficazmente los tweets, sino que también revela diferencias en el rendimiento basadas en la región. Esta información es crucial para entender las dinámicas de toxicidad en el discurso en línea, y aporta un nuevo enfoque al estudio de la toxicidad del lenguaje en redes sociales.

2. Interpretación de los resultados

Los resultados obtenidos en este proyecto revelan varios hallazgos importantes en la detección de lenguaje tóxico en español, tanto a nivel global como regional. A nivel global, los modelos basados en transformers, como DistilBERT, mostraron un rendimiento superior en la clasificación de lenguaje tóxico en comparación con los modelos tradicionales de machine

learning, como Logistic Regression, SVM, y Random Forest. DistilBERT demostró su capacidad para manejar la complejidad del lenguaje, logrando un f1-score de 0.75 en la clase tóxica, lo que lo convierte en la opción más efectiva para esta tarea. Sin embargo, los resultados también indicaron que, a pesar de su mejor rendimiento, el modelo sigue enfrentando desafíos para detectar tweets tóxicos con precisión, particularmente en contextos regionales.

A nivel regional, se encontró que los modelos tienden a funcionar mejor en Latinoamérica que en España, con SVM y Random Forest mostrando una mayor precisión y recall en la clasificación de comentarios tóxicos en Latinoamérica. Este comportamiento podría estar influido por las diferencias lingüísticas y culturales entre ambas regiones, lo que sugiere que los modelos, aunque efectivos, no son totalmente agnósticos al contexto geográfico o al dialecto. En España, se observó una disminución en el rendimiento de los modelos, especialmente en el recall de la clase tóxica, lo que indica una menor capacidad para capturar lenguaje ofensivo en esa región seguramente debido al desbalance del dataset analizado.

3. Limitaciones y trabajos futuros

Entre las limitaciones del estudio, se encuentra el tamaño del dataset, que, aunque adecuado, podría beneficiarse de una mayor diversidad en términos de contenido y contexto cultural para mejorar el rendimiento del modelo y su capacidad de generalización. Además, la distribución de los tweets entre las regiones no es completamente equilibrada, lo que puede haber influido en el rendimiento de los modelos.

Para futuras investigaciones, se sugiere explorar los siguientes aspectos:

1. **Aumento del dataset:** Ampliar el conjunto de datos con más tweets de diversas regiones y contextos o datos de otros temas relacionados con la toxicidad en el lenguaje en línea, como noticias o foros, para mejorar la capacidad del modelo de generalizar a otros contextos.
2. **Análisis de emojis y símbolos:** Incluir un análisis más profundo sobre cómo los emojis y otros símbolos afectan la percepción de toxicidad y cómo pueden ser incorporados en el modelo.
3. **Evaluación de sesgos en la clasificación:** Ampliar el análisis del sesgo por región (Latinoamérica y España) para evaluar si ciertos términos o expresiones en una región están asociados injustamente con mayor toxicidad. Investigar más a fondo cómo las características culturales y lingüísticas influyen en la clasificación de toxicidad.
4. **Modelos híbridos:** Investigar enfoques que combinan modelos de aprendizaje automático y técnicas basadas en reglas para abordar la identificación de toxicidad de manera más efectiva.
5. **Aplicación de modelos pre-entrenados más grandes:** Usar modelos más grandes y complejos como mBERT, XLM-RoBERTa, o BERTIN, que están pre-entrenados específicamente en textos en español. Estos modelos pueden mejorar significativamente el rendimiento en tweets escritos en español, especialmente en diferentes dialectos regionales.

6. **Modelos multilingües:** Probar modelos multilingües como XLM-RoBERTa o mBERT para evaluar si tienen mejor rendimiento al manejar tweets que mezclan diferentes dialectos del español o incluso combinan otros idiomas.
7. **Transferencia de aprendizaje:** Aplicar técnicas de transfer learning para aprovechar modelos ya entrenados en grandes corpus de datos y ajustar sus parámetros para mejorar la clasificación en español. Esto podría reducir el tiempo de entrenamiento y mejorar el rendimiento en nuevos datasets.
8. **Aplicaciones prácticas:** Examinar la implementación de este modelo en plataformas de redes sociales para moderación de contenido en tiempo real, así como su efectividad en entornos multilingües.
9. **Detección de sesgos sociales:** Explorar cómo los modelos actuales manejan sesgos sociales en el lenguaje, como el género, la raza o el origen étnico. Esto puede implicar la creación de conjuntos de datos específicos para evaluar cómo los modelos detectan toxicidad cuando está dirigida a grupos específicos.
10. **Detección de contexto y constructividad:** Ampliar el análisis para no solo clasificar la toxicidad, sino también la constructividad del comentario. Esto permitiría un análisis más matizado de los tweets, identificando comentarios que, aunque sean críticos, aportan valor a la conversación (López Úbeda et al. 2024).
11. **Evaluación continua:** Establecer un marco de evaluación continua para el modelo, de modo que pueda adaptarse a cambios en el lenguaje y en las dinámicas sociales a lo largo del tiempo.
12. **Explicabilidad de modelos:** Investigar técnicas de interpretabilidad de modelos para entender por qué un modelo clasifica un tweet como tóxico o no. Herramientas como LIME o SHAP pueden ayudar a identificar patrones en los datos que influyen en la decisión del modelo.

Estas líneas de investigación no solo contribuirían a mejorar la solución propuesta, sino que también permitirían expandir el conocimiento en el campo del análisis de contenido en redes sociales y su gestión.

REFERENCIAS

Bibliografía: las referencias clave para el proyecto

A continuación, se presenta una lista de las referencias bibliográficas relevantes que han sido citadas y consultadas a lo largo de este proyecto.

1. Álvarez-Carmona, M. A., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. En *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. (p. 75)
2. Arango Monnar, A., Perez, J., Poblete, B., Saldaña, M., & Proust, V. (2022). Resources for Multilingual Hate Speech Detection. En *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 122–130). Association for Computational Linguistics. <https://aclanthology.org/2022.woah-1.12/>
3. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. R., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. En *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)* (pp. 54-63). <https://doi.org/10.18653/v1/S19-2007>
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). Spanish Pre-trained BERT Model and Evaluation Data. *arXiv*. <https://arxiv.org/abs/2308.02976>
5. Das, A., Nandy, S., Saha, R., Das, S., & Saha, D. (2024). Analysis and Detection of Multilingual Hate Speech Using Transformer Based Deep Learning. *arXiv*. <https://arxiv.org/abs/2401.11021>
6. De la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. P., Romero, M. & Grandury, M. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *arXiv*. <https://arxiv.org/abs/2207.06814>
7. Devlin, J., Chang, M.V., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. <https://arxiv.org/abs/1810.04805>
8. Fersini, E., Rosso, P., & Anzovino, M. (2018). Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). *CEUR Workshop Proceedings*. <https://ceur-ws.org/Vol-2150/overview-AMI.pdf>
9. Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., & Vishwamitra, N. (2024). An Investigation of Large Language Models for Real-World Hate Speech Detection. *arXiv*. <https://arxiv.org/abs/2401.03346>
10. Hu, K., & Zhang, J. (2024). Toxicity Detection for Free. *arXiv*. <https://arxiv.org/abs/2405.18822>
11. Jahan, M. S. & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. <https://doi.org/10.1016/j.neucom.2023.126232>
12. Jurafski, D., & Martin, J. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/>

13. Koratana, A. & Hu, K. (2018). Toxic Speech Detection. En: *32nd Conference on Neural Information Processing Systems*.
<https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/15744362.pdf>
14. López Úbeda, P., Plaza-del-Arco, F., Díaz-Galiano, M., Martín-Valdivia, M. (2024) Toxicity in Spanish News Comments and its Relationship with Constructiveness. *Procesamiento del Lenguaje Natural*, v. 73.
<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6599>
15. Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
16. Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schaefer, J., Ranasinghe, T., Zampieri, M., Nandini, D., & Jaiswal, A. K. (2021). Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. *arXiv*. <https://arxiv.org/abs/2112.09301>
17. Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Multilingual hate speech detection: A semi-supervised generative adversarial approach. *Entropy*, 26(344).
<https://doi.org/10.3390/e26040344>
18. Mostafazadeh Davani, A., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11, 300–319. https://doi.org/10.1162/tacl_a_00550
19. Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, 19(21), 4654.
<https://doi.org/10.3390/s19214654>
20. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics.
<https://aclanthology.org/P19-1493/>
21. Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166, 114120.
<https://doi.org/10.1016/j.eswa.2020.114120>
22. Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). Multilingual HateCheck: Functional tests for multilingual hate speech detection models. En *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 154–169). Association for Computational Linguistics. <https://aclanthology.org/2022.woah-1.15/>
23. Sahin, E., Aydos, M., & Orhan, F. (2018). Spam/ham e-mail classification using machine learning methods based on bag of words technique. En *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4).
<https://doi.org/10.1109/SIU.2018.8404347>
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. *arXiv*. <https://arxiv.org/abs/1706.03762>
25. Zhang, J., Wu, Q., Xu, Y., Cao, C., Du, Z., & Psounis, K. (2023). Efficient Toxic Content Detection by Bootstrapping and Distilling Large Language Models. *arXiv*.
<https://arxiv.org/abs/2312.08303>

ANEXOS

Anexo 1: Tabla A1. Tabla de datasets relevados

Dataset	Url	Form ato	Origen	Regist ros	Descripción	Paper
Spanish Hate Speech Superset	https://huggingface.co/datasets/manueltonneau/spanish-hate-speech-superset	csv	Twitter	26421	Fusiona varios datasets en un solo archivo: hatEval, HaterNet, chileno, hascosva, HOMO-MEX. No accesible, hay que solicitarlo.	https://aclanthology.org/2024.woah-1.23/
hatEval, SemEval-2019 Task 5	https://github.com/msang/hateval/tree/master	tsv	Twitter	6600	Identifica lenguaje tóxico y agresivo	https://aclanthology.org/S19-2007.pdf
HaterNet	https://zenodo.org/records/2592149#.XmuNJahKg2w	txt	Twitter	6000	Tweets recolectados en España en 2017	https://www.mdpi.com/1424-8220/19/21/4654
chileno	https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/tree/master/Chilean%20dataset	csv	Twitter	9834	Tweets recolectados en Chile	https://aclanthology.org/2022.woah-1.pdf#page=136
HaSCoSva	https://gitlab.inria.fr/counter/HaSCoSva/-/blob/main/dataset/hascosva_2022_ano_nymized.tsv?ref_type=heads	tsv	Twitter	4000	Hate Speech Corpus with Spanish Variations (HaSCoSva-2022)	https://aclanthology.org/2023.vardial-1.1.pdf
HOMO-MEX	https://github.com/juanmvsa/HOMO-MEX/tree/main	csv	Twitter	862	Tweets recolectados en México	https://aclanthology.org/2023.woah-1.20/
AMI: Automatic Misogyny	https://amiibereval2018.wordpress.com/	xlsx	Twitter	4138	No accesible directamente,	https://ceur-ws.org/Vol-2150/overvi


Identification – IBEREVAL 2018	important-dates/data/				protegido	ew-AMI.pdf
MEX-A3T	https://github.com/lvonneMont/UDA_text_Mex-a3t/tree/main/text/data/mex-a3t	?	Twitter	11000	Tweets recolectados en México	https://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf
DACHS Hate speech and personal attack dataset in Spanish social media	https://zenodo.org/records/3520150	?	Twitter	37688	Tweets recolectados en España	https://www.sciencedirect.com/science/article/abs/pii/S2468696420300124
Multilingual Hate Speech Dataset for Fairness Evaluation	https://lrec2020.lrec-conf.org/en/shared-lrs/	tsv	Twitter	4831	Multilingüe	http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.180.pdf
NewsCom-TOX	https://detoxisiberlef.wixsite.com/website/corpus	zip	Comentarios periodicos online y foros	4357	Incluido en DETESTS	https://link.springer.com/article/10.1007/s10579-023-09711-x
OffendES	https://huggingface.co/datasets/fmplaza/offendes/tree/main	zip	Posts redes sociales (Twitter, Instagram, and YouTube)	30416	Correspondiente a la competición MeOffendES de IberLEF 2021	https://aclanthology.org/2021.ranlp-1.123.pdf
Jigsaw Multilingual Toxic Comment Classification	https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification	zip	Comentarios wikipedia	2480	Multilingüe, correspondiente a la competición de Kaggle	https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification
SpanishHopeEDI	https://codalab.lisn.upsaclay.fr/competitions/10215	zip	Twitter	1650	No accesible directamente, analiza el lenguaje esperanzador	https://ceur-ws.org/Vol-3625/paper6.pdf
DETESTS	https://huggingface.co/datasets/CLiC-UB/DETESTS-Dis	csv	Comentarios online	5629	Incluye los datasets NewsCom-Tox y el stereoHoax	https://ceur-ws.org/Vol-3202/overview.pdf
StereoHoax-ES	https://huggingface.co/datasets/CLiC-UB/DETESTS-Dis	csv	Twitter	5349	Incluido en Detests	http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6442

EXIST	https://nlp.uned.es/exist2021/	tsv	Twitter	5000	No disponible directamente, pero se envía por email bajo pedido	https://ceur-ws.org/Vol-3202/exist-aper8.pdf
Gender Bias in Spanish Tweets	https://www.kaggle.com/datasets/kevinmorgado/gender-bias-spanish	csv	Twitter	1914	Lenguaje sexista	https://www.kaggle.com/datasets/kevinmorgado/gender-bias-spanish
Paradetox	https://huggingface.co/datasets/textdetox/es_paradetox	tsv	Frases creadas ad hoc	565	Replica una misma frase con dos versiones: tóxica y neutral	https://aclanthology.org/2022.acl-long.469/
CONAN-SP	https://huggingface.co/datasets/SINAI/CONAN-SP	csv	Frases creadas ad hoc	238	Replica una misma frase con dos versiones: tóxica y neutral	https://rua.ua.es/dspace/bitstream/10045/137174/1/PLN_71_18.pdf
SHADES	https://huggingface.co/datasets/AnonymousSubmissionUser/shades	csv	Frases creadas ad hoc	649	Listado de estereotipos en varios idiomas	https://openreview.net/forum?id=zSwnz6BsDa&notId=zSwnz6BsDa

Anexo 2: Enlaces a Google Colab con el desarrollo del código

 GP_IEBS_Modelos.ipynb

<https://colab.research.google.com/drive/1swUHVy1V2Y7ZGFzxJtI7WRfBKLw0ZioN>

 GP_IEBS_Modelos_LenguajeTox.ipynb

<https://colab.research.google.com/drive/13yIBKDDnxzdYAjcHRI25sPvU5vsuk45>