Detector de Comentarios Tóxicos y Maliciosos en Redes Sociales

Diaz Lionel
COLEGIO UNIVERSITARIO IES - SIGLO 21
Procesamiento del Lenguaje Natural
Docente: Sufotinsky Alejandro

21 - 11 - 2024

COLEGIO UNIVERSITARIO IES Siglo 21

Introducción

La toxicidad en redes sociales representa una preocupación creciente en todo el mundo. Con el aumento de las plataformas de interacción virtual, ha surgido una proliferación de comportamientos agresivos, intimidación y contenido perjudicial que impacta tanto a usuarios como a creadores de contenido. Según un análisis de 2023, adolescentes y jóvenes son particularmente vulnerables a la presión social, el ciberacoso y el hostigamiento en línea, factores que pueden afectar negativamente su salud mental y bienestar emocional. Las tasas de depresión y ansiedad entre adolescentes han aumentado aproximadamente un 70% en los últimos 25 años, con un incremento adicional del 25% durante la pandemia (Consumer Notice, LLC, 15 de agosto de 2023).

Este entorno tóxico no solo afecta a los usuarios que consumen contenido, sino también a los creadores y moderadores de plataformas, quienes están expuestos a críticas abusivas y despectivas. En un artículo reciente, la American Psychological Association (APA) resaltó la necesidad urgente de mecanismos de protección para adolescentes en redes sociales, ya que muchos de ellos encuentran difícil desconectarse debido a la dependencia emocional que estos entornos pueden generar (Health Advisory on Social Media Use in Adolescence, 2023). La APA advierte que este contexto ha impulsado a legisladores y empresas a desarrollar medidas de mitigación, aunque los desafíos persisten por la naturaleza dinámica de las redes sociales (American Psychological Association, 2023).

Estudios también sugieren que la exposición continua a actitudes negativas en redes sociales puede contribuir al aumento de pensamientos y comportamientos autodestructivos entre adolescentes, quienes suelen compararse con los estándares idealizados que ven en estas plataformas. Un 54% de los adolescentes afirma pasar demasiado tiempo en redes, y muchos vinculan su bienestar emocional con la respuesta recibida en sus publicaciones. Esto refleja una preocupante correlación entre el tiempo de exposición y la salud mental, aunque otros factores como el contexto familiar y las experiencias personales también deben ser considerados (Consumer Notice, LLC, 15 de agosto de 2023).

Dada esta problemática, se ha identificado la necesidad de implementar herramientas automáticas que detecten y mitiguen niveles de agresión en redes sociales. Con este propósito, se han desarrollado sistemas que emplean técnicas de procesamiento del lenguaje

natural (NLP) y aprendizaje automático, los cuales permiten analizar el lenguaje en comentarios y detectar patrones de toxicidad. La combinación de NLP y modelos de machine learning ha demostrado ser eficaz en la creación de entornos más seguros, mejorando la experiencia del usuario al reducir la exposición a contenido dañino (van Aken, Risch, Krestel, & Löser, 2018; D'Sa, Illina, & Fohr, 2020).

Este trabajo de investigación tiene como objetivo desarrollar un programa capaz de detectar comentarios tóxicos en redes sociales mediante la aplicación de técnicas avanzadas de procesamiento del lenguaje natural, programación, machine learning e inteligencia artificial. El proyecto busca no solo identificar y clasificar comentarios ofensivos de manera eficaz, sino también contribuir a crear un entorno en línea más seguro y menos hostil, minimizando la exposición de los usuarios a contenido perjudicial.

Marco Teórico

1. Toxicidad en los Comentarios de Redes Sociales

La toxicidad en redes sociales es un problema persistente y complejo, especialmente en áreas de comentarios donde los usuarios expresan opiniones, críticas o reacciones a publicaciones. Estos comentarios tóxicos incluyen insultos, lenguaje ofensivo, discurso de odio y acoso, los cuales afectan tanto a los creadores de contenido como a los usuarios. En plataformas visuales como Instagram, la toxicidad frecuentemente se dirige hacia aspectos físicos o personales de los creadores, mientras que en redes más anónimas, como Twitter, permite una gran cantidad de comentarios polarizados y agresivos (Hinduja & Patchin, 2018).

Estudios recientes reflejan que un 41% de los adultos en los EE. UU. ha experimentado acoso en línea, y un 66% de estos incidentes han ocurrido en secciones de comentarios de redes sociales (Pew Research Center, 2017). Estos datos resaltan la necesidad de un sistema eficaz de moderación de comentarios, ya que en entornos con alta toxicidad la interacción positiva y la participación de los usuarios tienden a disminuir considerablemente (Geiger & Ford, 2021).

2. Procesamiento de Lenguaje Natural (NLP) y Técnicas Clave

El Procesamiento de Lenguaje Natural (NLP) es una rama de la inteligencia artificial que permite a las computadoras interpretar y analizar el lenguaje humano de manera precisa y útil. En el contexto de la moderación de redes sociales, el NLP ayuda a identificar patrones y relaciones en los textos para distinguir entre contenido inofensivo y comentarios tóxicos. Este proyecto aplica varias técnicas fundamentales de NLP para abordar el problema:

- **Tokenización**: La tokenización es el proceso de dividir el texto en unidades básicas, llamadas "tokens", que pueden ser palabras o frases. Esto facilita el análisis detallado de palabras y expresiones utilizadas en cada comentario. En el contexto de detección de toxicidad, esta técnica ayuda a descomponer los comentarios en elementos que el modelo puede interpretar para captar significados específicos.
- **Lematización y Stemming**: Estas técnicas de normalización reducen las palabras a su forma base o raíz, simplificando el análisis al reducir la variabilidad del lenguaje. La

lematización transforma las palabras a su forma canónica (e.g., "correr" y "corriendo" se convierten en "correr"), mientras que el stemming elimina sufijos para obtener las raíces. Ambas técnicas permiten que el modelo reconozca patrones clave en la toxicidad, independientemente de las variaciones gramaticales.

- **Embeddings**: Los embeddings son representaciones vectoriales que ubican palabras en un espacio semántico, de modo que la proximidad entre vectores indica similitud semántica. Técnicas como Word2Vec, GloVe y FastText permiten que el modelo identifique relaciones contextuales entre términos, lo que es crucial para la clasificación de toxicidad en comentarios. Word2Vec y GloVe capturan relaciones generales entre palabras, mientras que FastText permite captar palabras con errores ortográficos o nuevas variaciones, algo común en redes sociales.

Estas técnicas de NLP, en conjunto, permiten analizar el lenguaje de manera profunda y detectar señales sutiles de toxicidad, independientemente de cómo se estructuren o varíen los comentarios en el texto.

3. Representaciones de Texto y Embeddings

Los embeddings representan palabras en vectores numéricos en un espacio semántico, permitiendo que los modelos capten el contexto y las relaciones de significado entre palabras. En este proyecto se emplean dos técnicas de embeddings clave para abordar la detección de toxicidad en comentarios:

- **GloVe (Global Vectors for Word Representation)**: GloVe es una técnica que genera vectores de palabras basados en la coocurrencia de términos en un gran corpus de texto, lo que permite capturar relaciones semánticas generales entre palabras. Esto es útil en la clasificación de toxicidad, ya que ayuda al modelo a distinguir términos neutros de aquellos con connotaciones negativas o violentas.
- **FastText**: A diferencia de GloVe, FastText representa las palabras a nivel de subpalabras o caracteres, lo que permite captar palabras con errores ortográficos o variaciones no convencionales, comunes en comentarios de redes sociales. Esto facilita que el modelo detecte toxicidad incluso en textos donde se distorsiona la

ortografía para evadir filtros de contenido, mejorando la precisión en contextos de lenguaje informal.

El uso combinado de GloVe y FastText en este proyecto permite al sistema comprender tanto las relaciones semánticas generales como las variaciones específicas del lenguaje en redes sociales, maximizando la capacidad de detección de toxicidad en una amplia gama de estilos y estructuras lingüísticas.

4. Modelos de Machine Learning para Detección de Toxicidad

Para la detección de toxicidad en comentarios, existen múltiples enfoques en NLP que varían desde modelos básicos hasta arquitecturas avanzadas de *deep learning*. Los modelos seleccionados para este proyecto incluyen tanto métodos clásicos como técnicas modernas de redes neuronales:

- **Naive Bayes**: Este modelo probabilístico clasifica textos calculando la probabilidad de que una palabra pertenezca a una clase específica (por ejemplo, "tóxico" o "no tóxico"). Aunque asume independencia entre las palabras, su simplicidad lo hace efectivo para detectar insultos y palabras ofensivas en textos cortos. Naive Bayes es una opción básica pero útil para identificar términos individuales asociados a la toxicidad.
- **SVM (Support Vector Machines)**: Las SVM son modelos de clasificación que buscan el hiperplano que mejor separa las clases en un espacio de alta dimensionalidad, permitiendo clasificaciones precisas incluso en conjuntos de datos complejos. Sin embargo, este proyecto descartó su implementación final por la dificultad de SVM para captar relaciones contextuales profundas, algo esencial para comprender el tono de los comentarios.
- **LSTM (Long Short-Term Memory)**: Las redes LSTM, un tipo de red neuronal recurrente, son capaces de manejar secuencias de palabras en un texto, capturando dependencias contextuales a largo plazo. Este modelo es especialmente útil en la detección de toxicidad implícita y sarcasmo, ya que comprende el contexto completo en los comentarios. En este proyecto, LSTM es una pieza fundamental para analizar relaciones semánticas complejas en textos extensos.

- **BERT (Bidirectional Encoder Representations from Transformers)**: Aunque BERT ha demostrado ser muy eficaz en la clasificación de comentarios tóxicos debido a su capacidad de analizar el contexto bidireccional, su implementación en este proyecto fue inviable por las limitaciones computacionales. Para aplicaciones con mayores recursos, BERT sería una opción avanzada y efectiva para captar matices contextuales complejos.

5. Análisis de Sentimientos para Detección de Toxicidad

El análisis de sentimientos clasifica las emociones en textos (generalmente en categorías positivas, negativas o neutras) y es una herramienta útil para detectar tonos hostiles o agresivos en comentarios. En este proyecto, se utiliza el análisis de sentimientos como una capa adicional de detección de toxicidad, ya que permite identificar patrones emocionales que no necesariamente contienen insultos directos pero que reflejan un tono negativo o malicioso.

El modelo utilizado para esta tarea es *VADER* (*Valence Aware Dictionary for Sentiment Reasoning*), un analizador de sentimientos especializado en redes sociales, disponible en la biblioteca NLTK de Python. VADER asigna puntajes de negatividad y compuestos al texto, los cuales ayudan al sistema a captar tonos sutilmente negativos, complementando así la detección de toxicidad implícita en los comentarios.

6. Desafíos en la Detección de Toxicidad en Comentarios

La detección de toxicidad presenta varios desafíos técnicos y éticos. A continuación, se destacan los principales problemas:

- **Contexto y Ambigüedad**: Muchos comentarios pueden contener humor, sarcasmo o ironía, lo cual dificulta la detección de toxicidad. Los modelos deben interpretar el contexto para evitar clasificaciones erróneas. Por ejemplo, una frase sarcástica podría sonar negativa sin ser tóxica, mientras que comentarios más directos pueden ser fácilmente detectados.
- **Errores Ortográficos y Jerga**: Los comentarios en redes sociales a menudo contienen errores tipográficos o jerga específica que desafían a los modelos tradicionales. Técnicas como FastText, que puede captar sub-palabras, ayudan a

mitigar este problema al permitir que el sistema reconozca palabras con errores ortográficos o variaciones lingüísticas.

- **Sesgo y Equidad**: Los modelos de NLP pueden reflejar sesgos presentes en los datos de entrenamiento, afectando especialmente expresiones de ciertos grupos o dialectos que pueden ser clasificados erróneamente como tóxicos. Para reducir estos sesgos, es fundamental entrenar modelos en conjuntos de datos balanceados, así como monitorear constantemente los resultados para detectar posibles inconsistencias o sesgos en la clasificación.

7. Ensambles de Modelos en Detección de Toxicidad

En el contexto de la detección de toxicidad, los métodos de ensamble son útiles para mejorar la precisión y robustez del sistema. Al combinar varios modelos, un ensamble puede aprovechar las fortalezas individuales de cada uno y mitigar sus debilidades. En este proyecto, se implementa un ensamble de **Naive Bayes** y **LSTM**, junto con el análisis de sentimientos con *VADER*.

- **Naive Bayes**: Este modelo es eficiente para detectar palabras ofensivas o insultos explícitos en comentarios cortos, asignando un alto peso a términos específicos asociados con toxicidad. Su simplicidad lo convierte en una herramienta efectiva para identificar rápidamente lenguaje ofensivo directo.
- **LSTM**: Con una arquitectura avanzada, que incluye capas bidireccionales y de atención, LSTM captura contextos y relaciones semánticas profundas en comentarios largos y con toxicidad implícita. Esto permite al sistema reconocer matices en el lenguaje, mejorando la detección en comentarios complejos y sutiles.
- **Análisis de Sentimientos con VADER**: VADER complementa el ensamble al identificar el tono emocional de los comentarios, permitiendo captar toxicidad incluso cuando no hay palabras ofensivas explícitas, pero el tono general es negativo o malicioso.

Este ensamble, al combinar la detección precisa de palabras ofensivas con el análisis de contexto y tono, crea un sistema robusto y adaptable para la clasificación de comentarios en redes sociales. Al cubrir tanto toxicidad explícita como implícita, este enfoque optimiza la

precisión y mejora la eficacia del modelo en un entorno tan variado y dinámico como las redes sociales.

8. Estudios Previos en Detección de Toxicidad

Estudios recientes han evaluado distintos enfoques para la detección de toxicidad, arrojando resultados valiosos para identificar los modelos más efectivos:

- **Davidson et al. (2017)** probaron varios modelos de machine learning, incluidos Naive Bayes y SVM, para la detección de odio y ofensividad en Twitter. Concluyeron que los modelos basados en embeddings y redes neuronales obtuvieron mejores resultados frente a métodos tradicionales como Naive Bayes.
- **Zampieri et al. (2019)** evaluaron transformadores como BERT para detectar lenguaje abusivo en plataformas sociales, encontrando que BERT alcanzó una precisión significativamente superior en comparación con otros modelos, debido a su capacidad para capturar el contexto bidireccional en los comentarios.
- **Pavlopoulos et al. (2017)** propusieron un enfoque basado en redes neuronales recurrentes y embeddings preentrenados, logrando una alta precisión en la clasificación de comentarios de Wikipedia. Además, encontraron que la combinación de LSTM con embeddings como GloVe era particularmente eficaz en la detección de comentarios tóxicos en contextos diversos.
- **Competencia de Kaggle "Jigsaw Unintended Bias in Toxicity Classification" (2019**: En esta competencia, los participantes buscaron crear un modelo que no solo detectara toxicidad en los comentarios, sino que también minimizara los sesgos involuntarios que podían afectar a ciertos grupos o identidades. Los equipos mejor clasificados emplearon estrategias avanzadas, como modificaciones en el modelo BERT, para permitir el reponderado de muestras y así reducir sesgos. Combinaciones de modelos como BERT modificado con BiLSTM y técnicas de ajuste fino resultaron efectivas en la competencia, lo que ayudó a mejorar la precisión sin amplificar sesgos preexistentes en los datos de entrenamiento (como en comentarios de identidades no ofensivos que podrían etiquetarse erróneamente).

- Estos estudios subrayan que aunque los modelos transformadores son los más efectivos, existen alternativas, como LSTM con embeddings preentrenados, que pueden proporcionar buenos resultados en proyectos con recursos limitados.

Marco Metodológico

1. Introducción al Marco Metodológico

Este proyecto tiene como objetivo desarrollar un sistema capaz de detectar comentarios tóxicos en redes sociales mediante técnicas avanzadas de procesamiento de lenguaje natural (NLP) y modelos de aprendizaje automático. La toxicidad en los comentarios de plataformas sociales crea entornos hostiles que afectan la experiencia de los usuarios, tal como demuestran estudios previos (Hinduja & Patchin, 2018). Para enfrentar esta problemática, el proyecto adopta un enfoque metodológico mixto, que integra modelos de clasificación de texto con análisis de sentimientos. Esta combinación permite al sistema diferenciar con precisión los comentarios tóxicos de aquellos que no lo son, incluso cuando el lenguaje tóxico no es explícito.

Este marco metodológico detalla cada etapa del desarrollo, desde la recopilación y preprocesamiento de datos, hasta la creación e implementación del modelo de detección de toxicidad. La integración de técnicas de NLP y un ensamble de modelos garantiza una clasificación precisa y robusta, capaz de captar tanto la toxicidad explícita como la implícita en un entorno en línea.

2. Diseño del Proyecto

2.1 Descripción del Problema

La proliferación de contenido tóxico en redes sociales representa un desafío importante para los administradores y desarrolladores de estas plataformas. La toxicidad en los comentarios puede manifestarse en diversas formas, desde insultos directos y amenazas hasta expresiones más sutiles que promueven un ambiente degradante. Este proyecto se enfoca en la detección automática de comentarios tóxicos mediante modelos de aprendizaje automático, que pueden identificar patrones lingüísticos asociados con la toxicidad y clasificar los comentarios en función de su contenido.

2.2 Justificación de la Metodología Elegida

Para abordar la problemática, se ha seleccionado un enfoque metodológico basado en técnicas de NLP combinadas con algoritmos de aprendizaje automático. Esta elección permite analizar los textos de manera detallada y estructurada, optimizando la precisión en la clasificación de comentarios tóxicos. Las técnicas de preprocesamiento de texto, como la tokenización, lematización y el uso de embeddings, convierten el texto en representaciones numéricas que los modelos pueden interpretar y analizar con mayor efectividad.

La metodología incluye un ensamble de modelos: **Naive Bayes** y **LSTM**. Naive Bayes es eficaz para detectar insultos y palabras ofensivas en comentarios breves, mientras que LSTM, con su capacidad para captar dependencias contextuales, es adecuado para comentarios más largos o con toxicidad implícita. Además, se incorpora un análisis de sentimientos utilizando *VADER* (SentimentIntensityAnalyzer de la biblioteca NLTK), que complementa el ensamble permitiendo captar el tono emocional del comentario, lo cual es crucial para identificar toxicidad sutil o implícita. La combinación de estos enfoques permite maximizar la precisión en la detección de toxicidad y enfrentar de manera integral los diferentes tipos de lenguaje tóxico en redes sociales.

2.3 Enfoque Metodológico

El desarrollo del proyecto sigue una metodología experimental, que consiste en la implementación y evaluación de distintos modelos y técnicas de NLP para determinar cuál es más efectivo en la clasificación de comentarios tóxicos. Los pasos principales de este enfoque incluyen:

- **Recopilación de Datos**: La obtención de datos proviene de *datasets* especializados en clasificación de toxicidad, como el de la competencia de Kaggle “Toxic Comment Classification Challenge”. Este conjunto de datos contiene una amplia variedad de comentarios etiquetados según su nivel de toxicidad (e.g., insulto, amenaza), lo que proporciona una base sólida y diversa para entrenar y evaluar los modelos. Para simplificar el análisis, los comentarios se clasificaron de manera binaria en "tóxicos" y "no tóxicos", consolidando cualquier comentario marcado en una categoría de toxicidad como "tóxico".

- **Análisis Exploratorio de Datos (EDA)**: Se realizó un análisis preliminar para examinar la distribución de los datos, identificar características lingüísticas relevantes y comprender la proporción de comentarios tóxicos frente a los no tóxicos. Este análisis reveló un desequilibrio de clases, con una mayor cantidad de comentarios no tóxicos, lo que representa un desafío para el modelo. Para mitigar este problema, se implementaron ajustes de pesos en las clases y técnicas de sobremuestreo en la clase minoritaria (tóxicos) y submuestreo en la clase mayoritaria (no tóxicos).
- **Preprocesamiento de Texto**: Los datos de texto requieren una serie de transformaciones para que los modelos puedan procesarlos eficazmente. Para el modelo **Naive Bayes**, el preprocesamiento incluyó conversión a minúsculas, eliminación de enlaces, dígitos y puntuación, eliminación de *stopwords* y lematización para reducir la variabilidad en el vocabulario. En el caso del **LSTM**, se siguió un preprocesamiento orientado a preservar el contexto mediante tokenización y *padding* (relleno) de secuencias, lo cual permite que los textos se ajusten a una longitud uniforme.
- **Selección y Entrenamiento de Modelos**: La selección de modelos se basa en su capacidad para manejar distintos tipos de datos de texto. Naive Bayes y LSTM fueron elegidos como base para el ensamble, debido a sus características complementarias en la clasificación de comentarios cortos y en el análisis de dependencias contextuales. Además, se utilizaron embeddings como **GloVe** y **FastText**. GloVe captura relaciones generales de significado entre palabras, mientras que FastText permite al modelo reconocer variaciones ortográficas comunes en redes sociales, como errores de escritura o uso de jerga. Esta combinación de técnicas garantiza que el modelo capte tanto las relaciones semánticas amplias como las particularidades lingüísticas del entorno de redes sociales.

2.4 Objetivos Específicos del Proyecto

Los objetivos metodológicos específicos incluyen:

1. Desarrollar un modelo de detección de toxicidad con alta precisión, capaz de identificar tanto comentarios explícitamente ofensivos como aquellos con una toxicidad más sutil.

2. Implementar técnicas de NLP y análisis de sentimientos para proporcionar una base sólida en la clasificación de comentarios.
3. Integrar un ensamble de modelos que permita maximizar la eficiencia en la detección de toxicidad en redes sociales.
4. Evaluar el rendimiento del modelo mediante métricas específicas, como precisión, recall y F1-score, para garantizar su efectividad y viabilidad en una posible implementación en tiempo real.

Esta estructura metodológica proporciona una base sólida para el desarrollo de un sistema efectivo de detección de toxicidad, que se ajusta a las complejidades del lenguaje humano y a las particularidades de los comentarios en redes sociales.

3. Recopilación y Preprocesamiento de Datos

3.1 Selección de la Base de Datos

Para este proyecto, se evaluaron diversas bases de datos relacionadas con la clasificación de toxicidad en comentarios, con el objetivo de seleccionar la más adecuada para los objetivos del análisis. Tras esta investigación, se decidió utilizar la base de datos de la competencia de Kaggle "Toxic Comment Classification Challenge" debido a su riqueza y diversidad en comentarios etiquetados en múltiples categorías de toxicidad (como insulto, lenguaje ofensivo, y amenazas). Este conjunto de datos permite abordar la complejidad del lenguaje ofensivo en varios contextos.

Para los fines de este proyecto, se simplificaron las categorías en una clasificación binaria: “tóxicos” y “no tóxicos”. De este modo, cualquier comentario que esté marcado con al menos una categoría de toxicidad se clasifica como "tóxico", mientras que los demás se clasifican como "no tóxicos". Esta simplificación facilita la eficiencia y precisión del modelo, enfocándolo en una clasificación directa que es más adecuada para aplicaciones de moderación en tiempo real.

3.2 Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) se realizó para obtener una visión general de las características de los datos. Entre los hallazgos más importantes se encontró un marcado desbalance entre las clases, con aproximadamente 50,000 comentarios etiquetados como tóxicos y 106,000 como no tóxicos. Este desequilibrio representa un desafío para el modelo, ya que puede llevarlo a favorecer la clase mayoritaria, reduciendo así su sensibilidad hacia los comentarios tóxicos.

Para mitigar el problema del desbalance, se aplicaron dos estrategias complementarias:

1. **Ajuste de pesos en las clases**: Durante la fase de modelado, se asignaron mayores pesos a los comentarios tóxicos en la función de pérdida del modelo, incentivando así que el algoritmo preste más atención a la clase minoritaria.
2. **Sobremuestreo y submuestreo**: Para equilibrar la proporción de comentarios tóxicos y no tóxicos, se aplicaron técnicas de sobremuestreo en la clase minoritaria y submuestreo en la clase mayoritaria. Estas técnicas ayudan a crear un conjunto de datos más balanceado sin necesidad de modificar la estructura del modelo, mejorando la capacidad del sistema para detectar toxicidad sin sacrificar la diversidad de ejemplos.

3.3 Preprocesamiento de Texto

El preprocesamiento de texto fue personalizado para cada modelo, optimizando su rendimiento en la clasificación de comentarios. Los pasos principales para cada modelo fueron:

- **Para Naive Bayes**:
    - **Conversión a minúsculas**: Para uniformizar el texto y evitar duplicados causados por diferencias en el uso de mayúsculas.
    - **Eliminación de enlaces y dígitos**: Se eliminan todos los enlaces y números, ya que no aportan información relevante para la detección de toxicidad.
    - **Eliminación de puntuación y *stopwords***: Se eliminan caracteres de puntuación y palabras comunes sin valor semántico, facilitando un análisis centrado en los términos significativos.

    - **Lematización**: Reducción de palabras a su forma básica para mejorar la consistencia del vocabulario y facilitar el reconocimiento de términos clave relacionados con la toxicidad.
- **Para LSTM**:
    - **Conversión a minúsculas y eliminación de caracteres especiales**: Se transforman todos los caracteres a minúsculas y se eliminan aquellos que no sean letras para reducir el ruido en el texto.
    - **Eliminación de *stopwords* y lematización**: Igual que en Naive Bayes, se eliminan palabras comunes y se aplica lematización.
    - **Tokenización y *padding***: Se convierte cada palabra en un número basado en un vocabulario predefinido, y se aplica *padding* para que todas las secuencias tengan una longitud uniforme, lo cual permite a LSTM procesar comentarios de manera efectiva.
- **Embeddings con GloVe y FastText**: Para mejorar la representación semántica del texto, se emplean embeddings de GloVe y FastText. GloVe captura las relaciones generales entre palabras, permitiendo que el modelo distinga términos con connotaciones ofensivas o violentas. FastText, por su parte, ayuda a reconocer palabras con errores ortográficos o variantes de estilo comunes en redes sociales, lo que es esencial en un contexto donde los usuarios tienden a distorsionar la ortografía para evadir filtros de contenido.

Este preprocesamiento permite que cada modelo se enfoque en aspectos específicos del texto y aproveche al máximo sus características para la clasificación de toxicidad.

4. Implementación y Experimentación con Modelos

En esta sección se describe el proceso de implementación y experimentación de los diferentes modelos de aprendizaje automático, evaluando su rendimiento y determinando su viabilidad para la clasificación de comentarios tóxicos. Inicialmente, se probaron varios modelos de manera individual para evaluar su eficacia y decidir su inclusión en el ensamble final.

4.1 Implementación de Naive Bayes

Naive Bayes fue seleccionado como una de las primeras opciones debido a su simplicidad y eficiencia. Este modelo probabilístico es efectivo en textos cortos y se basa en la frecuencia de palabras individuales, permitiendo clasificar comentarios a partir de términos con alta probabilidad de toxicidad (como insultos y lenguaje ofensivo). Durante las pruebas, Naive Bayes mostró buenos resultados en la detección de comentarios explícitamente tóxicos, gracias a su capacidad para asignar probabilidades altas a palabras específicas. Sin embargo, presentó limitaciones en la clasificación de toxicidad implícita, donde el contexto juega un rol importante.

4.2 Implementación de Máquinas de Soporte Vectorial (SVM)

El modelo de Máquinas de Soporte Vectorial (SVM) fue evaluado para observar su capacidad en la clasificación binaria de comentarios. Este modelo es eficaz en problemas de clasificación de alta dimensionalidad, y se experimentó con varios tipos de núcleos (*kernels*). Aunque SVM mostró buenos resultados en métricas de entrenamiento, su rendimiento en comentarios nuevos fue inferior, especialmente en textos largos y con dependencias contextuales complejas. Por esta razón, SVM fue descartado en la configuración final del ensamble.

4.3 Implementación de Redes Neuronales LSTM

Las redes LSTM se utilizaron para modelar secuencias de texto y capturar dependencias contextuales a largo plazo, esenciales en comentarios complejos o con toxicidad implícita. Se desarrolló una arquitectura personalizada para maximizar su rendimiento, que incluyó:

- **Capa bidireccional**: Esta capa permite que LSTM procese el texto en ambas direcciones, lo cual mejora la comprensión del contexto completo de cada comentario, especialmente en aquellos donde el significado cambia en función de la estructura de la frase.
- **Capa de atención**: La atención se emplea para dar mayor peso a palabras o frases que son más relevantes para detectar toxicidad, lo cual permite identificar comentarios que pueden tener un tono ofensivo sin palabras explícitas.

- **Embeddings de GloVe y FastText**: La combinación de estos embeddings permite a LSTM captar tanto el contexto semántico general como las variaciones lingüísticas, como errores ortográficos comunes en redes sociales.

Este modelo demostró ser efectivo para identificar patrones complejos de toxicidad, especialmente en comentarios largos y con matices contextuales.

4.4 Ensamblaje de Modelos: Naive Bayes, LSTM y Análisis de Sentimientos

Tras evaluar el rendimiento de cada modelo, se implementó un ensamble que integra Naive Bayes, LSTM y el análisis de sentimientos con *VADER*. Este ensamble combina las fortalezas de cada enfoque para mejorar la precisión en la detección de toxicidad:

- **Naive Bayes**: Detecta con rapidez insultos y palabras ofensivas en comentarios breves, asignando un peso elevado a términos específicos.
- **LSTM**: Con su capacidad para analizar contextos y secuencias largas, detecta toxicidad implícita en comentarios complejos y con dependencia contextual.
- **Análisis de Sentimientos (VADER)**: Este análisis complementa el ensamble evaluando el tono emocional de los comentarios, permitiendo captar actitudes negativas o maliciosas en casos donde el lenguaje ofensivo es más sutil.

La combinación de estos tres enfoques optimiza la clasificación de toxicidad al cubrir una amplia gama de patrones lingüísticos y matices emocionales, lo que hace que el sistema sea robusto y efectivo en entornos de redes sociales.

4.5 Implementación del Modelo Final: Flujo de Trabajo del Ensamblaje

El ensamblaje final combina las predicciones de **Naive Bayes**, **LSTM** y el análisis de sentimientos con *VADER* para maximizar la precisión y robustez del sistema. A continuación, se describe el flujo de trabajo implementado para calcular la probabilidad final de toxicidad y clasificar los comentarios:

1. **Preprocesamiento del Comentario**:
    - El comentario se somete a una limpieza inicial (conversión a minúsculas, eliminación de caracteres no alfabéticos, entre otros) y luego se adapta a los

requisitos de cada modelo. Para Naive Bayes, se genera una representación basada en TF-IDF, mientras que para LSTM, se realiza tokenización y *padding* para garantizar que las secuencias tengan una longitud uniforme.

2. **Cálculo de Probabilidad de Toxicidad con Naive Bayes**:
    - Utilizando la representación TF-IDF, el modelo asigna una probabilidad de toxicidad basada en la presencia y peso de palabras clave asociadas a la toxicidad. Este cálculo es especialmente útil para identificar insultos y lenguaje explícitamente ofensivo.
3. **Cálculo de Probabilidad de Toxicidad con LSTM**:
    - El comentario procesado es evaluado por la red LSTM, que analiza el contexto completo y las relaciones entre palabras para detectar toxicidad implícita. La probabilidad generada refleja la interpretación contextual del texto, siendo especialmente eficaz en comentarios largos o complejos.
4. **Análisis de Sentimientos con VADER**:
    - El análisis de sentimientos calcula un puntaje de negatividad y un valor compuesto para el comentario. Si el valor compuesto indica un tono altamente negativo, esta probabilidad se ajusta para reflejar un mayor nivel de toxicidad, complementando así las predicciones de los otros modelos.
5. **Combinación de Probabilidades**:
    - Las probabilidades generadas por cada modelo se combinan utilizando una fórmula ponderada. Las ponderaciones fueron definidas de acuerdo con la fortaleza de cada modelo en diferentes escenarios:
        - **Naive Bayes**: 35%
        - **LSTM**: 40%
        - **Análisis de Sentimientos (VADER)**: 25%
    - Esta combinación permite que el sistema aproveche las fortalezas de cada componente, garantizando una evaluación equilibrada que considera tanto toxicidad explícita como implícita y el tono emocional del comentario.
6. **Clasificación Final**:
    - Basándose en la probabilidad ponderada final, los comentarios se clasifican en tres categorías:
        - **No Tóxico**: Probabilidad ≤ 0.45

- **Posible Tóxico**: Probabilidad entre 0.46 y 0.65
- **Tóxico**: Probabilidad > 0.65

5 Resultados

5.1 Resultados de Naive Bayes

El modelo **Naive Bayes**, diseñado para clasificar comentarios tóxicos basándose en la probabilidad de palabras individuales, demostró un rendimiento sólido en textos cortos con lenguaje explícito. Las métricas clave fueron:

- **Entrenamiento**:
  - **Precisión Global**: 90.68%
  - **Clase 0 (No Tóxico)**: Precisión: 93%, Recall: 88%, F1-Score: 90%
  - **Clase 1 (Tóxico)**: Precisión: 89%, Recall: 93%, F1-Score: 91%
- **Pruebas con Nuevos Comentarios**:
  - Precisión General: 70% (promedio).
  - **Fortalezas**: Detección precisa de comentarios con insultos explícitos o palabras ofensivas frecuentes.
  - **Limitaciones**: Incapacidad para interpretar el contexto; por ejemplo, comentarios sarcásticos o con toxicidad implícita a menudo fueron clasificados erróneamente como "No Tóxicos".

El rendimiento en comentarios nuevos mostró que Naive Bayes depende en gran medida de la presencia de palabras clave, lo que lo hace menos efectivo en textos con variaciones lingüísticas o donde el tono general es negativo sin insultos explícitos.

5.2 Resultados de SVM

El modelo **SVM (Support Vector Machines)** fue evaluado como una alternativa para la clasificación binaria, destacándose inicialmente en el conjunto de entrenamiento:

- **Entrenamiento**:
  - **Precisión Global**: 98.34%
  - **Clase 0 (No Tóxico)**: Precisión: 100%, Recall: 97%, F1-Score: 98%
  - **Clase 1 (Tóxico)**: Precisión: 97%, Recall: 100%, F1-Score: 98%

- **Pruebas con Nuevos Comentarios**:
    - Precisión General: Menos del 50%.
    - **Fortalezas**: Separación efectiva de clases en textos bien estructurados y con características distinguibles.
    - **Limitaciones**: Dificultad para manejar comentarios con estructuras complejas o matices contextuales. Por ejemplo, comentarios con sarcasmo o frases donde el contexto determina la toxicidad fueron clasificados incorrectamente.

El bajo rendimiento en datos nuevos sugiere que SVM es menos adaptable a las variaciones de lenguaje presentes en redes sociales, lo que lo hace inadecuado para este proyecto.

5.3 Resultados de LSTM

El modelo **LSTM**, diseñado para manejar secuencias largas y dependencias contextuales, mostró el mejor rendimiento entre los modelos individuales:

- **Entrenamiento**:
    - **Precisión Global**: 93% (alcanzada en 10 épocas).
    - **Loss Mínima**: 0.17.
- **Pruebas con Nuevos Comentarios**:
    - Precisión General: 83%.
    - **Fortalezas**: Excelente en la detección de toxicidad implícita y comentarios largos con contexto complejo. Por ejemplo, frases donde el sarcasmo o el tono general sugieren toxicidad fueron clasificados correctamente.
    - **Limitaciones**: Menor eficacia en textos muy breves o con palabras explícitas aisladas, donde el contexto no aporta información suficiente.

LSTM superó a Naive Bayes y SVM en su capacidad para comprender patrones más complejos en comentarios que dependen de la estructura o el tono para transmitir toxicidad.

5.4 Resultados del Ensamblaje Final

El ensamblaje final combinó las predicciones de **Naive Bayes**, **LSTM** y el análisis de sentimientos con *VADER*, logrando un rendimiento significativamente superior al de los modelos individuales:

- **Pruebas con Nuevos Comentarios**:
    - Precisión Global: 91%.
    - Clasificaciones Correctas:
        - **No Tóxico**: 89% de precisión.
        - **Posible Tóxico**: 85% de precisión.
        - **Tóxico**: 94% de precisión.
- **Fortalezas del Ensamblaje**:
    - **Naive Bayes**: Detección rápida y precisa de insultos explícitos.
    - **LSTM**: Análisis profundo de contextos complejos y dependencias a largo plazo.
    - **Análisis de Sentimientos**: Mejora en la detección de toxicidad implícita en comentarios con tono emocional negativo.

El ensamblaje mostró un rendimiento notable en comentarios nuevos gracias a la combinación de técnicas complementarias.

Conclusiones

El desarrollo de este proyecto permitió explorar y validar un enfoque robusto para la detección de comentarios tóxicos en redes sociales mediante la combinación de técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. Los resultados obtenidos demuestran que un modelo ensamblado, compuesto por **Naive Bayes**, **LSTM** y el análisis de sentimientos con *VADER*, ofrece un rendimiento significativamente superior al de los modelos individuales, alcanzando una precisión global del **91%** en pruebas con nuevos comentarios.

1. Análisis de Resultados y Logros

1.1 Fortalezas del Ensamblaje:

- El modelo ensamblado logra abordar tanto la toxicidad explícita como la implícita. **Naive Bayes** destaca en la identificación de insultos y palabras ofensivas explícitas, mientras que **LSTM** se especializa en interpretar el contexto y los matices semánticos en comentarios más complejos. El análisis de sentimientos con *VADER* complementa estas capacidades al captar el tono emocional de los textos, mejorando la detección en casos ambiguos.
- Este enfoque equilibrado permitió clasificar correctamente más del **90% de los comentarios** en escenarios variados, desde insultos directos hasta expresiones sutiles o sarcásticas.

1.2 Cumplimiento de los Objetivos:

- El proyecto alcanzó los objetivos iniciales al desarrollar un sistema eficaz para clasificar comentarios tóxicos, contribuyendo a la creación de entornos digitales más seguros. Además, el uso de técnicas de NLP y modelos de aprendizaje automático permitió abordar los desafíos inherentes al lenguaje informal y dinámico de las redes sociales.

1.3 Relevancia Social:

- La implementación de herramientas como esta puede tener un impacto significativo en la moderación de contenido en redes sociales, al reducir la exposición de los usuarios a mensajes dañinos. Este modelo también puede servir como una herramienta de apoyo para los equipos de moderación, optimizando el proceso de filtrado de comentarios y permitiendo un manejo más eficiente de la toxicidad en línea.

2. Limitaciones del Proyecto

Aunque el ensamblaje demostró ser efectivo, el proyecto enfrenta algunas limitaciones:

- **Restricciones Computacionales**:
  - La imposibilidad de implementar BERT debido a sus altos requerimientos computacionales limitó la exploración de arquitecturas más avanzadas. Esto sugiere la necesidad de recursos adicionales para futuros desarrollos.
- **Dependencia del Conjunto de Datos**:
  - El modelo se entrenó y probó utilizando un conjunto de datos específico que, aunque diverso, puede no representar completamente la variabilidad del lenguaje y toxicidad en redes sociales.
- **Lenguaje y Contexto Cultural**:
  - Las diferencias lingüísticas y culturales en la percepción de toxicidad no fueron abordadas en esta iteración del modelo, lo que limita su aplicación global sin ajustes adicionales.

3. Futuras Líneas de Trabajo

3.1 Ampliación del Conjunto de Datos:

- Incorporar datos provenientes de múltiples plataformas y contextos lingüísticos, lo que permitirá al modelo adaptarse mejor a la variabilidad del lenguaje y sus usos en diferentes regiones y culturas.

3.2 Exploración de Modelos Avanzados:

- Implementar modelos más ligeros basados en transformadores, como **DistilBERT**, que ofrecen resultados comparables a BERT con menores requerimientos computacionales.

3.3 Validación en Entornos Reales:

- Probar el modelo en plataformas reales de redes sociales para evaluar su rendimiento frente a datos no controlados y obtener retroalimentación directa de su efectividad en moderación automatizada.

3.4 Mitigación de Sesgos:

- Diseñar estrategias para identificar y reducir sesgos inherentes en los datos de entrenamiento, garantizando que el modelo sea justo en la clasificación de comentarios provenientes de diversos grupos sociales y culturales.

3.5 Optimización para Tiempo Real:

- Optimizar la arquitectura del modelo ensamblado para que pueda procesar grandes volúmenes de comentarios en tiempo real, asegurando su viabilidad para aplicaciones prácticas.

Conclusión General

Este proyecto demuestra la viabilidad de integrar modelos de aprendizaje automático y técnicas de NLP para abordar un problema crítico en redes sociales: la toxicidad en los comentarios. A través de un modelo ensamblado, se logró una clasificación precisa y eficiente, sentando las bases para futuras implementaciones en entornos reales. Aunque existen desafíos técnicos y éticos por superar, los resultados obtenidos subrayan el potencial de esta tecnología para promover interacciones digitales más seguras y saludables.

Referencias

Doctor Mark C. Howell Jr. - “Social Media & Mental Health Statistics” - Consumer Notice (2023) https://www.consumernotice.org/personal-injury/social-media-harm/statistics/

Protecting Teens on Social Media - APA (2023) https://www.apa.org/monitor/2023/09/protecting-teens-on-social-media

Health advisory on social media use in adolescence - APA (2023) https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification. Challenges for Toxic Comment Classification: An In-Depth Error Analysis

D'Sa, A. G., Illina, I., & Fohr, D. (2020). Towards non-toxic landscapes: Automatic toxic comment detection using DNN. Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN

Hinduja & Patchin - Cyberbullying.org (2018). https://cyberbullying.org/ - https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2018.pdf

Jigsaw Unintended Bias in Toxicity Classification - Kaggle (2019). https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Moon, Jihyung et al. (2023). "Analyzing Norm Violations in Live-Stream Chat". arXiv (ar5iv)(ACL Anthology).

Kanfoud, MR et al. (2022). "Automated Tool for Toxic Comments Identification on Live Streaming YouTube". SpringerLink(SpringerLink).

Salminen, J et al. (2020). "Developing an Online Hate Classifier for Multiple Social Media Platforms". SpringerLink(SpringerLink).

Nandakumar, R et al. (2022). "Sentiment Analysis on Student Feedback Using NLP". IEEE Xplore(SpringerLink).

NerdTitanTV (2023). "How Twitch Streamers Handle Toxicity, Hate, Harassment, and Trolls". NerdTitanTV(NerdTitan TV).