ISSN 1870-4069

Metodología para la detección de misoginia en comentarios de Youtube

Belém Priego Sánchez, Kevin Melgoza Rodriguez

Universidad Autónoma Metropolitana Unidad Azcapotzalco,
División de Ciencias Básicas e Ingeniería,
México

{abps, al2222001696}@azc.uam.mx

Resumen. Este artículo presenta una metodología para la detección de contenido misógino en comentarios de YouTube para el idioma español, aplicando técnicas de procesamiento de lenguaje natural. El corpus utilizado, se recopiló de videos relacionados con una controversia pública, y se aplicó un léxico de discurso de odio alineado con la definición de misoginia de la Real Academia Española. Se crearon varios modelos de aprendizaje automático supervisado: árboles de decisión, Naïve Bayes, así como máquinas de vectores de soporte, y un modelo básico de aprendizaje profundo basado en BiLSTM para lograr una clasificación binaria. El modelo de aprendizaje profundo proporcionó los mejores resultados, para la detección de misoginia, con un 94 % de precisión durante el entrenamiento basado en una división 80/20. Este resultado se debe al enfoque bidireccional, que ha facilitado la evaluación contextual en el corpus y posteriormente ha enriquecido a los resultados finales. La metodología propuesta, muestra que el uso de PLN y el modelado computacional del lenguaje permiten un análisis del discurso para la detección de odio relacionado con el género en espacios digitales.

Palabras clave: Aprendizaje automático, misoginia, comentarios en español.

Methodology for the Detection of Misogyny in YouTube Comments

Abstract. This paper presents a methodology for detecting misogynistic content in Spanish-language YouTube comments using natural language processing techniques. The corpus was collected from videos related to a public controversy, and a hate speech lexicon aligned with the definition of misogyny established by the Real Academia Española was applied. Several supervised machine learning models were developed, including decision trees, Naïve Bayes, and support vector machines, as well as a deep learning model based on BiLSTM for binary classification. The deep learning model achieved the best performance in misogyny detection, reaching 94 % accuracy during training using an 80/20 data split. This

result is attributed to the bidirectional architecture, which enabled improved contextual analysis within the corpus and consequently enhanced the final results. The proposed methodology demonstrates that the use of NLP and computational language modeling facilitates discourse analysis for the detection of gender-related hate speech in digital environments.

Keywords: Machine learning, misogyny, comments in Spanish.

1. Introducción

La creciente producción de contenido generado por usuarios en plataformas digitales como YouTube ha intensificado el interés por desarrollar métodos automáticos capaces de analizar grandes volúmenes de texto de forma eficiente. El procesamiento del lenguaje natural (PLN) permite transformar comentarios no estructurados en información analizable, permitiendo identificar patrones discursivos y fenómenos sociales relevantes a gran escala [6].

Uno de estos fenómenos es la misoginia, definida por la Real Academia Española (RAE) como la "aversión hacia las mujeres" [11], que se manifiesta en el discurso a través de expresiones de rechazo, hostilidad o desprecio. En los entornos digitales, estas manifestaciones pueden adoptar formas explícitas, como insultos directos, o aparecer de manera implícita mediante ironía, estereotipos o lenguaje aparentemente neutro, lo que dificulta su detección automática.

Diversos estudios han abordado la detección automática de discurso misógino utilizando técnicas de aprendizaje automático y modelos de lenguaje produndo. En [7] y [4] se han propuesto enfoques basados en clasificación supervisada para identificar contenido ofensivo dirigido a mujeres en redes sociales. Asimismo, investigaciones más recientes han incorporado modelos basados en transformadores, como BERT, que han demostrado un rendimiento superior en tareas de clasificación de texto al capturar mejor el contexto semántico [13].

En el ámbito hispanohablante, también se han desarrollado iniciativas orientadas a la detección de discurso de odio y misoginia, destacando la creación de corpus anotados y la adaptación de modelos multilingües para mejorar la precisión en español [8]. Además, algunas investigaciones señalan que una proporción significativa de usuarios, especialmente jóvenes, han estado expuestas a este tipo de discurso y percibe un aumento del lenguaje hostil en línea [3]. De igual manera, se ha documentado que ciertas comunidades digitales concentran niveles particularmente elevados de expresiones misóginas, lo que refuerza la necesidad de mecanismos sistemáticos para su análisis y clasificación [10].

Al respecto, el aumento del discurso hostil en línea y la concentración de expresiones misóginas en ciertas comunidades digitales evidencian la magnitud del problema y la necesidad de herramientas que permitan abordarlo. La identificación manual de este tipo de contenido resulta insuficiente frente al volumen de datos generado en plataformas digitales, en consecuencia es necesario el desarrollo de técnicas que automaticen y facilitan tanto la detección, como el análisis.

Por lo que, en este artículo se presenta una metodología para la detección automática de misoginia en comentarios de YouTube en español, basada en técnicas de PLN y modelos de aprendizaje automático y aprendizaje profundo. El objetivo principal es mostrar la existencia del fenómeno y describir las etapas que permitan recopilar, procesar, representar y clasificar comentarios de manera estructurada, sentando así una propuesta para el análisis automatizado del discurso misógino en plataformas digitales.

1.1. Estado del arte

La detección automática de misoginia permite identificar, analizar y moderar contenido que promueve o reproduce actitudes de odio, discriminación o violencia contra las mujeres en entornos digitales. La investigación, de esta área de estudio e investigación, ha cambiado en los últimos años a partir de enfoques clásicos de aprendizaje automático hacia modelos más avanzados de procesamiento de lenguaje natural.

En trabajos iniciales se emplearon algoritmos como Máquinas de Soporte Vectorial (SVM), Naive Bayes, regresión logística y clasificadores en conjunto, así como redes neuronales recurrentes (LSTM) y convolucionales (CNN). Estos métodos han mostrado un buen desempeño en tareas de clasificación binaria, aunque presentan limitaciones para capturar el contexto y las expresiones implícitas de misoginia [15]. Con el desarrollo de modelos de lenguaje pre-entrenados basados en transformadores, como BERT y RoBERTa, se han reportado mejoras significativas en la detección de discurso agresivo y misógino. En la competencia TRAC-2, diversos equipos utilizaron configuraciones multitarea basadas en BERT, logrando incrementos en métricas como el F1-score frente a enfoques tradicionales [13].

Asimismo, estudios recientes han explorado arquitecturas híbridas y modelos multilingües para escenarios bilingües, obteniendo mejores resultados en precisión y recall [9]. La mayoría de los trabajos se han centrado en textos en inglés o en contextos multilingües con lenguas ampliamente representadas. Sin embargo, en los últimos años han surgido investigaciones orientadas al español, tanto en la creación de corpus anotados como en el diseño de modelos específicos para la detección automática de misoginia; algunos estudios han propuesto el uso de modelos adaptados y la combinación de múltiples fuentes de características para abordar la escasez de datos en este idioma [1]. Además, se han presentado nuevos recursos lingüísticos para lenguas menos estudiadas, como el desarrollo de un corpus específico para la detección de misoginia en gallego, lo que evidencia un interés creciente por ampliar estos estudios a otros contextos lingüísticos y culturales [2]. Estos trabajos muestran un avance progresivo en el uso de técnicas de PLN para la detección de misoginia, así como una transición desde enfoques clásicos hacia modelos basados en representaciones pre-entrenadas.

Por ello, mantener una idea clara de los distintos tipos de misoginia, como el anti-feminista, el deshumanizador, la violencia domestica, los estereotipos, la interseccional, la falta de autonomía, el falocentrismo, la violencia sexual, la sexualizacion, la transfobia/homofobia y la trivializacion [14]. De esta manera,

se puede visualizar el reto que supone generar un modelo y metodología que vaya más allá de una clasificación binaria —misógino o no misógino—, que permita revisar y matizar las diferentes manifestaciones de misoginia presentes en la interacción textual de los entornos digitales.

2. Detección de misoginia

En esta sección se describe la metodología propuesta, detallando cada una de las fases involucradas en la construcción del corpus, el preprocesamiento lingüístico y la implementación de modelos de clasificación orientados a la identificación de comentarios misóginos de YouTube. El enfoque se basa en un flujo de trabajo práctico de clasificación de texto, desde la recopilación de los comentarios hasta el entrenamiento y evaluación de distintos modelos de aprendizaje automático y aprendizaje profundo.

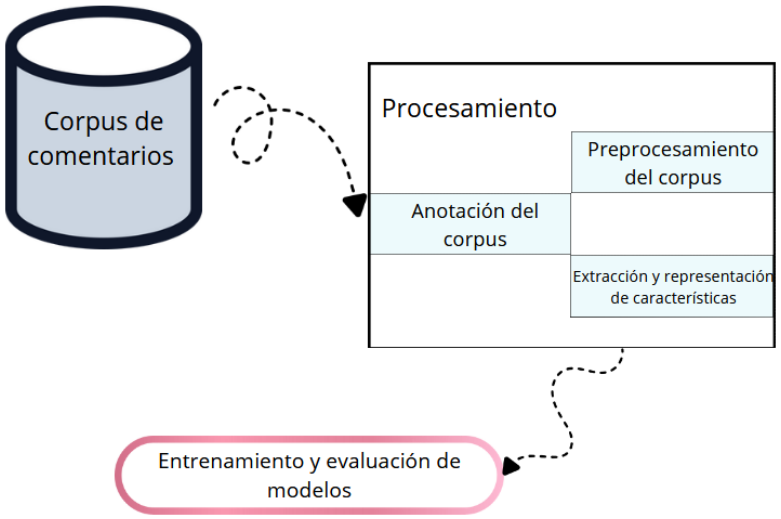

Fig. 1. Esquema general de la metodología propuesta

2.1. Metodología propuesta

La metodología se estructuró como una tarea de clasificación de texto compuesta por varias etapas consecutivas. El proceso comienza con la obtención de comentarios generados por la comunidad usuaria y finaliza con la evaluación de modelos entrenados para identificar contenido misógino. De forma general, el flujo de trabajo se divide en cuatro fases: (1) recopilación de comentarios en español desde YouTube, (2) preparación y anotación del conjunto de datos, (3)

preprocesamiento y representación del texto, y (4) entrenamiento y evaluación de modelos de clasificación. Este diseño permite mantener coherencia entre los datos utilizados, las características extraídas y los modelos implementados.

2.2. Construcción del conjunto de datos

En esta sección se describe el proceso seguido para la construcción del conjunto de datos utilizado en este artículo, así como los criterios empleados para su anotación y preparación. Se detalla la estrategia de recolección de comentarios desde la plataforma de YouTube, seguida del enfoque adoptado para la identificación automática de contenido misógino mediante un lexicón especializado y, finalmente, se explica el procedimiento aplicado para balancear el corpus, con el objetivo de garantizar su idoneidad para el entrenamiento y evaluación de los modelos de aprendizaje automático.

Recolección

Esta etapa consistió en reunir un conjunto comentarios asociados con una controversia pública: un futbolista generó una fuerte controversia en julio de 2025 debido a comentarios publicados en sus redes sociales, considerados sexistas y machistas por gran parte de la opinión pública, medios de comunicación y usuarios.

Los comentarios se recopilaron, del período de finales de julio a principios de septiembre de 2025, utilizando la API de Datos de YouTube; esto se llevó a cabo a partir de los identificadores (códigos únicos) de videos seleccionados asociados a la controversia del caso de estudio.

Se extrajeron los comentarios asociados a cada video junto con metadatos básicos, como el nombre del canal, el autor, el texto original del comentario y la fecha de publicación. Este procedimiento permitió construir un conjunto inicial de 56,352 comentarios en español, que sirven como base para el análisis posterior.

Anotación

La anotación del conjunto de datos se realizó mediante un enfoque automático basado en un lexicón de términos de discurso de odio en español [12], tomando como referencia la definición de misoginia proporcionada por la RAE para mantener consistencia conceptual. Se estableció un umbral de siete coincidencias léxicas[1], una coincidencia ocurre cada vez que una palabra o frase del comentario coincide exactamente o de manera normalizada con alguna entrada del lexicón asociado a expresiones misóginas, de modo que un comentario es clasificado como misógino únicamente si contiene al menos siete términos del lexicón, lo que

[1] Las coincidencias léxicas hacen referencia a la presencia de términos o expresiones específicas, previamente definidas en el lexicón de discurso de odio, dentro del texto de cada comentario.

permite reducir falsos positivos y asegurar que la etiqueta asignada corresponda a una presencia significativa de lenguaje ofensivo dirigido hacia las mujeres. La asignación de las etiquetas binarias a los comentarios, está basada en:

- 1 = misógino,
- 0 = no misógino.

Balanceo Debido al fuerte desbalance entre clases observado, tras la anotación inicial, y con el fin de evitar sesgar el aprendizaje de los modelos de clasificación (favoreciendo la clase mayoritaria y reduciendo su capacidad para identificar correctamente instancias de la clase minoritaria); el conjunto de datos fue reducido y balanceado a un total de 10,000 comentarios, logrando una distribución alrededor del 50 % para cada clase (misógino y no misógino), adecuada para el entrenamiento de los modelos. Adicionalmente, la reducción del tamaño del corpus permitió mantener un conjunto manejable computacionalmente sin comprometer la diversidad de ejemplos, facilitó un entrenamiento más eficiente y una evaluación más equitativa del desempeño de los modelos.

2.3. Preprocesamiento lingüístico

En esta etapa se describe el proceso de preprocesamiento y transformación de los datos textuales, así como la construcción y evaluación de los modelos de clasificación utilizados en este artículo. Inicialmente, se aplicaron técnicas de limpieza y normalización para reducir el ruido presente en los comentarios y estandarizar el formato. Posteriormente, se llevaron a cabo procesos lingüísticos como la tokenización y lematización, junto con la generación de características que permiten representar el texto de manera adecuada para su análisis computacional. Estas representaciones sirvieron como base para el entrenamiento de modelos de aprendizaje automático y de aprendizaje profundo, cuya configuración, entrenamiento y evaluación se detallan en los apartados siguientes.

Limpieza y normalización Los comentarios se sometieron a un preprocesamiento estándar con el objetivo de reducir ruido y homogeneizar el texto. Este proceso incluyó la conversión a minúsculas, la eliminación de URL, menciones, signos de puntuación, números y espacios innecesarios. Adicionalmente, se filtraron los comentarios para conservar únicamente aquellos con una longitud aproximada de entre 25 y 300 palabras, evitando así sesgos derivados de textos demasiado cortos o excesivamente largos.

Tokenización, lematización y generación de características Como parte del preprocesamiento lingüístico, se generó una segunda hoja de trabajo en la que cada comentario fue transformado a una representación estructurada basada en unidades lingüísticas. En primer lugar, se aplicó la tokenización, dividiendo cada texto en palabras o tokens individuales, lo que permite analizar de forma granular el contenido del comentario. Posteriormente, se realizó la lematización, proceso

mediante el cual cada token se redujo a su forma canónica o lema (por ejemplo: "mujeres", "mujer" → "mujer"), con el fin de unificar variantes morfológicas y mejorar la consistencia del análisis.

Adicionalmente, se calcularon diversas estadísticas léxicas, tales como la longitud del comentario (número de palabras), el número de tokens únicos, la riqueza léxica (relación entre palabras únicas y totales) y la frecuencia de aparición de términos relevantes. Asimismo, se incorporaron descriptores lingüísticos complementarios, como la proporción de palabras ofensivas según el lexicón utilizado, la presencia de pronombres, adjetivos calificativos, así como indicadores de intensidad discursiva (por ejemplo, uso de mayúsculas o repeticiones). Esta información enriquecida permitió capturar no solo el contenido semántico del texto, sino también patrones estructurales y estilísticos, facilitando la generación de características más informativas para los modelos de clasificación.

Representación de características Para los modelos de aprendizaje automático tradicionales, los textos se representaron mediante vectores de características TF-IDF (*Term Frequency – Inverse Document Frequency*) utilizando unigramas y bigramas. En el caso del modelo de aprendizaje profundo, los comentarios fueron tokenizados y preprocesados en secuencias numéricas para asegurar que todas las secuencias de texto (comentarios) en el conjunto de datos tengan la misma longitud (*padded sequences*) con un tamaño de vocabulario fijo.

2.4. Modelos de clasificación

En esta fase se aborda la construcción y entrenamiento de los modelos de clasificación empleados para identificar comentarios misóginos. Se exploraron dos enfoques complementarios: primero, modelos de aprendizaje automático clásicos, que permiten establecer una línea base sólida y evaluar patrones discriminativos a partir de representaciones TF-IDF del texto; y segundo, un modelo de aprendizaje profundo basado en una arquitectura BiLSTM, diseñado para capturar relaciones contextuales y dependencias semánticas en los comentarios.

Esta combinación de enfoques permite comparar métodos tradicionales y modernos, evaluando su capacidad para detectar tanto expresiones explícitas como implícitas de misoginia en el corpus.

Modelos de aprendizaje automático Se entrenaron tres algoritmos clásicos ampliamente utilizados en tareas de clasificación de texto, los cuales son: Árboles de Decisión, Máquina de Vectores de Soporte (SVM) y Naïve Bayes, debido a su efectividad probada en problemas de categorización de documentos, su capacidad para manejar conjuntos de datos de tamaño moderado y su interpretabilidad. Estos modelos se entrenaron utilizando las representaciones TF-IDF del corpus previamente procesado.

Los Árboles de Decisión permiten identificar reglas explícitas que diferencian clases, SVM es robusto frente a datos de alta dimensionalidad y dispersos, como los generados por TF-IDF, y Naïve Bayes, a pesar de su simplicidad, ofrece buen

desempeño en tareas de clasificación de texto debido a su modelo probabilístico que asume independencia entre términos, facilitando una primera línea base confiable para comparación con métodos más complejos.

Modelo de aprendizaje profundo Además de los modelos clásicos, se implementó un modelo de red neuronal basado en una arquitectura *Bidirectional LSTM (BiLSTM)*, con el objetivo de capturar dependencias contextuales en los comentarios. Este tipo de arquitectura resulta especialmente útil para identificar expresiones implícitas o sutiles de misoginia.

El modelo se construyó sobre una capa de *embedding* inicializada con vectores preentrenados de GloVe obtenidos del *Spanish Billion Word Corpus* [5]. Se utilizó un vocabulario de 40,000 palabras y vectores de 300 dimensiones, los cuales se mantuvieron congelados durante el entrenamiento para preservar la información semántica original.

La arquitectura final estuvo compuesta por:

- una capa de embedding preentrenada con secuencias acolchadas de longitud máxima 300;
- una capa BiLSTM con 128 unidades, *dropout* y *recurrent dropout* de 0.2;
- una capa densa de 64 unidades con activación ReLU y *dropout* de 0.4;
- una capa de salida con activación sigmoide para clasificación binaria.

2.5. Entrenamiento y evaluación del modelo

Los modelos, anteriormente descritos, se entrenaron utilizando una división estratificada del conjunto de datos en 80 % para entrenamiento y 20 % para prueba. Durante el entrenamiento se emplearon los *callbacks EarlyStopping* y *ReduceLROnPlateau* con el fin de evitar sobreajuste y optimizar el proceso. Se utilizó la función de pérdida *binary crossentropy* y el optimizador *Adam*.

La división estratificada del conjunto de datos en 80 % para entrenamiento y 20 % para prueba garantiza que la proporción de clases (misógino y no misógino) se mantenga constante en ambos subconjuntos, evitando sesgos en la evaluación y asegurando que los modelos aprendan patrones representativos de cada clase.

El uso de los *callbacks EarlyStopping* y *ReduceLROnPlateau* permite controlar el entrenamiento de manera eficiente: el primero detiene el entrenamiento cuando la métrica de validación deja de mejorar, evitando sobreajuste; el segundo ajusta automáticamente la tasa de aprendizaje cuando la mejora se estanca, optimizando la convergencia.

La función de pérdida *binary crossentropy* es adecuada para tareas de clasificación binaria, ya que mide la diferencia entre las probabilidades predichas y las verdaderas etiquetas, mientras que el optimizador *Adam* combina eficiencia computacional con adaptabilidad en la actualización de los pesos, facilitando un entrenamiento estable y rápido de los modelos de aprendizaje profundo.

La evaluación se realizó sobre el conjunto de prueba mediante métricas estándar como precisión, recall y F1-score, además del análisis de la matriz de confusión. Finalmente, tanto el modelo entrenado como el tokenizador fueron

almacenados para su posible reutilización, asegurando la reproducibilidad del experimento y facilitando análisis posteriores.

3. Resultados

El objetivo principal de esta sección es comparar el desempeño de enfoques clásicos de aprendizaje automático frente a un modelo de aprendizaje profundo, analizando su capacidad para distinguir entre comentarios misóginos y no misóginos.

La evaluación se realizó sobre un conjunto de prueba compuesto por 2,000 comentarios (20 % del corpus balanceado) previamente procesados y anotados, esto dado que permite medir de manera objetiva el desempeño de los modelos sobre datos no vistos durante el entrenamiento, asegurando que las métricas reflejen su capacidad real de generalización. Al mantener la proporción de clases consistente con el conjunto original, se evita que el sesgo de distribución afecte la evaluación, y se garantiza que tanto comentarios misóginos como no misóginos estén representados adecuadamente para un análisis comparativo confiable.

Para cada modelo de clasificación se calcularon métricas estándares de evaluación, considerando tanto el desempeño global como el comportamiento individual en cada clase.

La Tabla 1 resume los resultados obtenidos del rendimiento obtenido por los diferentes modelos de clasificación evaluados sobre el conjunto de prueba. Se presentan cuatro modelos: Árbol de Decisión, SVM Lineal, Naïve Bayes y BiLSTM Bidireccional. En ella se reportan métricas generales, como *accuracy* y *F1 Measure*, junto con métricas específicas por clase, lo que permite un análisis más detallado del comportamiento de los modelos en la detección de comentarios misóginos (clase 1) y no misóginos (clase 0).

Tabla 1. Métricas de rendimiento de los modelos de clasificación.

Modelo	**Accuracy**	**F1 Measure**	**Precisión (1)**	**Recall (1)**	**F1 (1)**	**Precisión (0)**	**Recall (0)**	**F1 (0)**
Árbol de Decisión	0.853	0.852	0.846	0.869	0.857	0.859	0.836	0.847
SVM Lineal	0.936	0.938	0.914	0.966	0.939	0.962	0.905	0.933
Naive Bayes	0.713	0.706	0.674	0.847	0.751	0.783	0.573	0.662
BiLSTM Bidireccional	0.945	0.945	0.929	0.967	0.947	0.964	0.922	0.943

A partir de los resultados, de la Tabla 1, se observa que el BiLSTM bidireccional obtuvo el mejor desempeño general, con un $accuracy = 0{,}945$ y métricas *F1* altas para ambas clases, lo que indica que puede capturar tanto patrones explícitos como contextuales de misoginia.

El SVM Lineal también muestra un rendimiento competitivo, especialmente en la detección de comentarios misóginos ($F1 = 0{,}939$ para la clase 1), mientras que el Árbol de Decisión obtiene resultados sólidos pero ligeramente inferiores. En contraste, Naïve Bayes presenta el desempeño más bajo, particularmente en

la clase negativa, lo que sugiere que suposiciones de independencia de términos limitan su efectividad en este corpus.

En general, los resultados permiten comparar cómo cada enfoque maneja la detección de comentarios misóginos y no misóginos, destacando la ventaja de las redes neuronales bidireccionales frente a los modelos tradicionales en tareas de clasificación de texto contextual.

3.1. Análisis y discusión de resultados

Los modelos clásicos presentan comportamientos contrastantes: mientras que el Árbol de Decisión logra un desempeño relativamente equilibrado entre ambas clases, su capacidad de generalización es limitada en comparación con otros métodos. En particular, se observan errores en la clasificación de comentarios no misóginos, lo que sugiere una sensibilidad moderada al ruido presente en el texto.

El modelo SVM Lineal destaca entre los enfoques tradicionales, alcanzando valores elevados en las métricas globales y un mejor balance entre *precisión* y *recall*, especialmente en la clase misógina. Este comportamiento indica una mayor capacidad para separar ambas clases a partir de representaciones TF-IDF, incluso en presencia de lenguaje informal y variado.

Por el contrario, Naïve Bayes obtiene el rendimiento más bajo entre los modelos evaluados. Sus resultados reflejan dificultades para capturar dependencias más complejas en el texto, lo que afecta especialmente la clasificación de la clase no misógina. Este comportamiento es consistente con las limitaciones conocidas de este modelo en tareas donde el contexto juega un papel relevante.

El mejor desempeño global corresponde al modelo BiLSTM bidireccional con embeddings preentrenados en español. Este modelo supera a los enfoques clásicos tanto en métricas generales como en las métricas por clase, mostrando un comportamiento más estable y equilibrado. La arquitectura bidireccional permite aprovechar el contexto completo de los comentarios, mientras que el uso de *embeddings preentrenados* contribuye a una mejor representación semántica del lenguaje, lo que resulta especialmente útil para identificar expresiones sutiles o implícitas de misoginia.

Estos resultados muestran que los modelos de aprendizaje profundo, cuando se apoyan en representaciones lingüísticas adecuadas, ofrecen ventajas claras frente a enfoques tradicionales en la tarea de detección automática de misoginia en comentarios en español.

4. Conclusiones y perspectivas

En esta sección se presentan las conclusiones derivadas del análisis realizado con la detección de comentarios misóginos en YouTube. Se resumen los resultados obtenidos por los distintos modelos de clasificación, se discuten las implicaciones de los enfoques empleados y se reflexiona sobre las limitaciones del estudio. También, se señalan posibles líneas de trabajo futuras que podrían mejorar la

eficacia de los sistemas de detección y enriquecer el análisis del discurso en plataformas digitales.

La detección de misoginia es un desafío relevante en el análisis de contenido en plataformas digitales, dado que este tipo de discurso puede manifestarse de formas explícitas e implícitas, variando en tono, contexto y estructura lingüística. Identificar comentarios misóginos de manera automática requiere reconocer palabras ofensivas, comprender el contexto y las sutilezas del lenguaje, por lo que combinar técnicas de procesamiento de texto, representación semántica y modelos de aprendizaje automático o profundo es imprescindible.

Los resultados evidencian que la elección del modelo de clasificación tiene un impacto significativo en la detección de comentarios misóginos en YouTube. Los modelos clásicos de aprendizaje automático, como Árbol de Decisión, SVM Lineal y Naïve Bayes, presentan comportamientos contrastantes: mientras que el Árbol de Decisión logra un desempeño equilibrado ($accuracy = 0{,}85$) pero limitado en generalización, el SVM Lineal alcanza un alto rendimiento ($accuracy = 0{,}93$), especialmente en la identificación de comentarios misóginos, gracias a su capacidad de separar clases en espacios de alta dimensionalidad. Por su parte, Naïve Bayes demuestra ser insuficiente ($accuracy = 0{,}71$) para capturar dependencias contextuales complejas, lo que reduce su efectividad en la clasificación precisa de comentarios no misóginos.

El modelo BiLSTM bidireccional, al incorporar embeddings preentrenados en español y aprovechar el contexto completo de los comentarios, supera de manera consistente a los enfoques tradicionales, logrando métricas elevadas y equilibradas para ambas clases ($accuracy = 0{,}94$). Lo que confirma que los modelos de aprendizaje profundo son adecuados para tareas de detección de misoginia que requieren interpretar tanto patrones explícitos como implícitos en el lenguaje.

Estos resultados destacan la importancia de seleccionar arquitecturas y representaciones de texto que capturen la semántica y el contexto del lenguaje. Este artículo plantea una base para futuras investigaciones y aplicaciones en la moderación automática de contenidos en plataformas digitales en español. Los resultados del enfoque de esta metodología son prometedores, el uso primario de las mínimas instancias de entrenamiento ayudaron a simplificar la tarea de la formación de un modelo de detección de misoginia funcional.

Existen distintos factores que hacen que un modelo se sesgue tanto el conjunto de datos, como el algoritmo de entrenamiento o simplemente los parámetros con que fue configurada; sin embargo, el enfoque propuesto toma un corpus asociado a una controversia en redes sociales con alta presencia de discurso misógino, el cual resultó ser clave para los resultados de este estudio.

A lo largo del desarrollo de esta investigación se identificaron varios aspectos críticos: uno de los más relevantes fue la distribución del conjunto de datos, ya que el desbalance entre clases puede afectar directamente el comportamiento de los modelos y su capacidad de generalizar.

A pesar de los avances obtenidos, todavía existen importantes áreas de oportunidad en este ámbito. Por mencionar algunas, el conjunto de datos es princi-

palmente la fuente de aprendizaje, pero lo fundamental radica en probar con distintos algoritmos, modelos preentrenados, o simplemente con modelos ya definidos e inspeccionar los resultados en busca del más óptimo. Una simple regla binaria de si o no, aun no es suficiente para identificar coincidencias sutiles de misoginia o violencia simbólica. De esta manera, la integración de modelos más robustos es la clave para llegar a una resolución, haciendo especial énfasis en las formas sutiles de la misoginia.

Referencias

1. Aldana-Bobadilla, E., Molina-Villegas, A., Montelongo-Padilla, Y., Lopez-Arevalo, I., Sordia, O.: A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers. Applied Sciences 11, no. 21, 10467 (2021). https://doi.org/10.3390/app112110467
2. Álvarez-Crespo, L. M., Castro, L. M.: A Galician Corpus for Misogyny Detection Online. Proceedings of the 16th International Conference on Computational Processing of Portuguese, Vol. 1, pp. 22–31, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics (2024). https://aclanthology.org/2024.propor-1.3/
3. Amnesty International UK. Toxic tech: new polling exposes widespread online misogyny driving Gen Z away from social media (2025). https://www.amnesty.org.uk/press-releases/toxic-tech-new-polling-exposes-widespread-online-misogyny-driving-gen-z-away-social
4. Anzovino, M., Fersini, E., Rosso, P.: Automatic Identification and Classification of Misogynistic Language on Twitter. Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB 2018), pp. 57–64. Springer, (2018). https://doi.org/10.1007/978-3-319-91947-8_6
5. Cardellino, C.: Spanish Billion Words Corpus and Embeddings, (2019). https://crscardellino.github.io/SBWCE/
6. Datasumi: Challenges and Future in Natural Language Processing, 2025. https://en.datasumi.com/challenges-and-future-in-natural-language-processing
7. Fersini, E., Rosso, P., Anzovino, M.: Overview of the Task on Automatic Misogyny Identification at IberEval 2018. Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Sevilla, España, pp. 214–228. CEUR Workshop Proceedings, vol. 2150 (2018). https://ceur-ws.org/Vol-2150/overview-AMI.pdf
8. García-Díaz, J. A., Cánovas-García, M., Palacios, R. C., Valencia-García, R.: Detecting misogyny in Spanish tweets: An approach based on linguistic features and word embeddings. Future Generation Computer Systems, 114, 506–518 (2021). https://doi.org/10.1016/j.future.2020.08.032
9. Hashmi, E., Yayilgan, S.Y., Yamin, M.M. et al.: Enhancing misogyny detection in bilingual texts using explainable AI and multilingual fine-tuned transformers. Complex Intell. Syst. 11, 39 (2025). https://doi.org/10.1007/s40747-024-01655-1
10. Kara-Yakoubian, M.: Incel forums reveal persistent, widespread misogyny regardless of user engagement. (2024). https://www.psypost.org/incel-forums-reveal-persistent-widespread-misogyny-regardless-of-user-engagement/
11. Real Academia Española y Asociación de Academias de la Lengua Española: Diccionario panhispánico de dudas (DPD), entrada “misoginia”: “aversión a las mujeres”. https://www.rae.es/dpd/misoginia

12. Said-Hung, E., Römer Pieretti, M., Montero-Díaz, J., De Lucas Vicente, A., Martínez Torres, J.: Hate Speech Library in Spanish / Librería de odio en Español (Version 2) [Dataset] (2023). https://doi.org/10.6084/m9.figshare.22383643.v2
13. Samghabadi, N., Patwa, P., PYKL, P., Mukherjee, P., Das, A., Solorio, T.: Aggression and Misogyny Detection using BERT: A Multi-Task Approach. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 126–131, Marseille, France, European Language Resources Association (ELRA) (2020). https://aclanthology.org/2020.trac-1.20/
14. Sheppard, B., Richter, A. Cohen, A., Smith, E. A., Kneese, T., Pelletier, C., Baldini, I., Dong, Y.: Subtle Misogyny Detection and Mitigation: An Expert-Annotated Dataset, arXiv:2311.09443 (2023). https://arxiv.org/pdf/2311.09443
15. Shushkevich, E., Cardiff, J.: Automatic Misogyny Detection in Social Media: A Survey. Computación y Sistemas, 23(4), 1159–1164 (2021). https://doi.org/10.13053/cys-23-4-3299