https://doi.org/10.37815/rte.v36n1.1088
Artículos originales

Sistema de Aprendizaje Automático para la Detección y Análisis de Contenido Sexista en la Música Urbana

A Machine Learning System for Detection and Analysis of Sexist Content in Urban Music

Dany Pianchiche-Añapa[1] https://orcid.org/0009-0001-6875-8120,
Pablo Pico-Valencia[1] https://orcid.org/0000-0003-3518-3313, Juan A. Holgado-Terriza[2]
https://orcid.org/0000-0002-8031-1276

[1]*Pontificia Universidad Católica del Ecuador*, Esmeraldas, Ecuador
dany.pianchiche@pucese.edu.ec, pablo.pico@pucese.edu.ec

[2]*Universidad de Granada*, Granada, España
jholgado@ugr.ec

Enviado: 2023/10/28
Aceptado: 2024/06/27
Publicado: 2024/06/30

Resumen

En este artículo se presentan los aspectos relacionados con la creación de un clasificador automático destinado a evaluar y categorizar el nivel de sexismo presente en las letras de canciones del género musical urbano. El sistema de clasificación asigna las letras a tres categorías distintas: "A", indicando contenido apto para audiencias de todas las edades; "B", señalando contenido que requiere supervisión de adultos; y "C", representando material orientado a adultos. El clasificador se implementó en Python aplicando los algoritmos Naïve Bayes, vecinos más cercanos, árbol de decisión, máquina de vectores de soporte y regresión logística. Para el proceso de entrenamiento de los modelos se creó un set de datos compuesto por 479 observaciones, dividido en un 75% para entrenamiento y un 25% para pruebas. El set de datos de entrenamiento abarcó tanto expresiones con connotaciones sexistas como aquellas que carecen de ellas. El clasificador que alcanzó el más alto grado de precisión fue el modelo basado en el algoritmo de regresión logística con un 77% de precisión. Con el fin de facilitar la explotación del clasificador en entornos de producción, se integró el modelo con una interfaz gráfica de usuario que facilita la usabilidad del sistema a los potenciales beneficiarios.

Palabras clave: Aprendizaje supervisado, clasificador, sexismo, música urbana, inteligencia artificial.

Sumario: Introducción, Materiales y Métodos, Resultados y Discusión, Conclusiones.

Como citar: Pianchiche-Añapa, D., Pico-Valencia, P. & Holgado-Terriza, J. (2024). Sistema de Aprendizaje Automático para la Detección y Análisis de Contenido Sexista en la Música Urbana. *Revista Tecnológica - Espol, 36(1), 68-80.* https://rte.espol.edu.ec/index.php/tecnologica/article/view/1088

Abstract

This paper presents aspects related to the creation of an automatic classifier designed to evaluate and categorize the level of sexism present in the lyrics of songs of the urban music genre. The classification system assigns lyrics to three different categories: "A", indicating content suitable for audiences of all ages; "B", indicating content requiring adult supervision; and "C", representing adult-oriented material. The classifier was implemented in Python by applying the following algorithms: Naïve Bayes, nearest neighbours, decision tree, support vector machine and logistic regression. For the model training process, a dataset composed of 479 observations was created, divided into 75% for training and 25% for testing. The training dataset included both expressions with sexist connotations and those without. The classifier that achieved the highest degree of accuracy was the model based on the logistic regression algorithm with 77% accuracy. In order to facilitate the exploitation of the classifier in production environments, the model was integrated with a graphical user interface that facilitates the usability of the system for potential beneficiaries.

Keywords: Supervised learning, classifier, sexism, urban music, artificial intelligence.

Introducción

El sexismo, definido como la discriminación o los estereotipos basados en el género, continúa siendo un problema arraigado en la sociedad contemporánea. Según el Diccionario de Oxford, el sexismo es la "discriminación o estereotipos contra las mujeres, sobre la base del sexo" (OED, 1866). De manera similar, la Real Academia de la lengua Española, lo define como la "discriminación de personas por motivos de sexo" (RAE, 2023). En la actualidad, la discriminación y la desigualdad de género persisten y se han intensificado en el ámbito tecnológico (Dhrodia, 2017). Internet, en cierta medida, perpetúa las diferencias de género y las actitudes sexistas en el mundo moderno en el que predominan las nuevas tecnologías de la información y las comunicaciones.

En el contexto del ciberespacio, las redes sociales constituyen un medio para generar comportamientos no aceptados socialmente y compartir contenidos sexistas y de muchas otras índoles tales como de acoso escolar y de incitación al odio (Mesiti & Yeo, 2023). En vista de que las redes sociales pueden ser accedidas por cualquier tipo de usuario que tenga un dispositivo inteligente con conexión a Internet y conocimiento de tecnologías, permite que niños puedan acceder a ciertos contenidos de manera indiscriminada. En este sentido, los padres sienten preocupación de que sus hijos vean y escuchen cosas que no son adecuadas para su edad. Aunque las redes sociales son empleadas por los niños y adolescentes principalmente para comunicarse; también las usan ampliamente para ver videos digitales y escuchar canciones, muchas de ellas enmarcadas en el género urbano (reguetón) —uno de los géneros musicales que más se escucha en la actualidad en Iberoamérica y que tiene la capacidad de transformar las normas sociales que regulan el comportamiento de los adolescentes (Penagos Rojas, 2012).

Las nuevas formas de crear este género musical han llamado mucho la atención en sus contenidos porque en sus letras incorporan terminologías o corpus sexistas que ofenden a las mujeres (Piñón Lora & Pulido Moreno, 2020) o a ciertas comunidades, como es el caso de la comunidad intersexual (LGBTI), pudiendo llegar a ser desagradables para la sociedad en general y, especialmente, para los niños. Por citar un ejemplo de estos cantantes se tiene a Maluma que se hizo controversial por la canción *Four Babys,* disponible en su canal de YouTube. Muchos han criticado esta canción porque determinan que denigra a la mujer y que contiene bastantes términos indecorosos que cualquier padre en su sano juicio no permitiría que sus hijos lo escuchen. Otro cantante que se ha hecho conocer por las letras de sus canciones,

pero no precisamente por ser educativas, es Bad Bunny. En la mayoría de sus éxitos como “Soy Peor” y “Te Boté”, disponibles en su canal de YouTube, hablan de sexo, drogas, dinero, mujeres y fiestas. Estas canciones son actualmente un gran problema para las familias que desean inculcar a sus hijos valores morales como el respeto, la igualdad y la equidad, de manera que a futuro sean personas de bien que aporten positivamente en la sociedad (Piñón Lora & Pulido Moreno, 2020).

La Inteligencia Artificial (IA), a través del Procesamiento del Lenguaje Natural (PLN), minería de datos y el aprendizaje profundo, permiten detectar contenidos específicos —cibercrimen, acoso escolar, homofobia, agresividad— en diferentes formatos digitales como audio, imagen, video y texto (Arce-García & Menéndez-Mendéndez, 2023; Castañeda Muñoz, 2019; Lepe, 2021). La minería de texto es un conjunto de "estrategias de recuperación de información no tradicional" (Ghosh et al., 2012), orientadas a reducir el esfuerzo requerido de los usuarios para obtener información útil de grandes fuentes de datos de texto computarizados. A través de la minería de texto se han propuesto varios estudios encaminados a realizar análisis de sentimientos. El análisis de sentimientos aplica la técnica de aprendizaje automático para, a partir de un conjunto de datos, crear un clasificador que aprenda bajo el paradigma del aprendizaje supervisado y así realizar el análisis de sentimientos (positivos, negativos, neutros), tomando en consideración datos nuevos distintos a los usados en el entrenamiento del clasificador. De esta manera, se ha empleado esta técnica para evaluar opiniones de las publicaciones en Twitter, ahora X (Cedeño-Moreno & Vargas, 2020).

El análisis de sentimientos y discursos de odio en X es un tema de gran interés y relevancia en la actualidad, debido al creciente uso de esta plataforma para expresar opiniones, emociones y actitudes sobre diversos temas. Para realizarlo, se han empleado diversas técnicas de minería de texto, aprendizaje automático y aprendizaje profundo, las cuales se han aplicado a diferentes fuentes de datos y redes sociales ampliamente usadas a nivel mundial como es el caso de X, Facebook, Instagram, TikTok, entre otras.

Uno de los aspectos más importantes para el análisis de sentimientos y discursos de odio es la selección y extracción de características relevantes de los textos, las cuales pueden ser desde palabras individuales hasta n-gramas, frecuencias, vectores semánticos, entre otras. En este sentido, Rasel et al. (2018) propusieron un método para filtrar los comentarios cibernéticos agresivos en tres categorías: discurso de odio, discurso ofensivo o ninguno de los dos anteriores. Para ello, utilizaron técnicas de minería de textos como n-grama y TF-IDF (Frecuencia de Ocurrencia del Término en la Colección de Documentos), las que sirvieron para preparar las entradas para los clasificadores de aprendizaje automático como: bosque aleatorio, regresión logística y máquina de soporte vectorial. Los autores reportaron una precisión del 93% con el bosque aleatorio, superando a los otros dos métodos. Asimismo, encontraron que la combinación de TF-IDF, Análisis Semántico Latente (LSA) y análisis de similitud coseno produjo un vector de características optimizado para la clasificación.

Otro aspecto relevante para el análisis de sentimientos y discursos de odio es el uso de modelos de aprendizaje profundo, los que pueden capturar relaciones complejas y no lineales entre las características y las etiquetas. En este sentido, Zhang et al. (2018) utilizaron una Red Convolucional (CNN) con una Unidad Recurrente Cerrada (GRU), combinada con incrustaciones de palabras. Con este modelo, lograron detectar los discursos de odio en Twitter con una alta precisión medida por la métrica F1. Los autores demostraron que la combinación del modelo CNN y GRU mejoró empíricamente la precisión de la clasificación en comparación con otros modelos basados en Redes Neuronales Recurrentes (RNN) o CNN.

Además del uso de modelos de aprendizaje profundo, también se han explorado otras técnicas de aprendizaje automático para el análisis de sentimientos y discursos de odio. Por ejemplo, Jiang & Suzuki (2019) realizaron varias comparaciones usando distintas proporciones de datos extraídos de Twitter con diferentes técnicas al mismo tiempo. Como hallazgo, el estudio evidenció que el aprendizaje automático puede llegar a tener un buen rendimiento cuando los datos son pequeños; sin embargo, cuando los datos son voluminosos se obtuvieron hallazgos más relevantes al implementar el aprendizaje profundo. También, se afirmó que el uso del modelo RNN obtuvo mejores resultados en comparación con otros métodos que se utilizaron, como es el caso de la regresión logística y las máquinas de soporte vectorial. RNN, implementada a partir de GRU, reportó el 96,56 % de precisión en las pruebas realizadas con el set de datos B usado en el experimento.

Finalmente, también se han aplicado técnicas de PLN para el análisis de sentimientos y discursos de odio en redes sociales. Por ejemplo, Back, B. H., & Ha (2019) propusieron un sistema para extraer información de sentimientos humanos de grandes cantidades de datos no estructurados, usando el algoritmo Naïve Bayes y técnicas del PLN para preprocesar el contenido de la información. Los resultados experimentales mostraron que el método proporcionó una precisión del 63.5%, que fue superior al método basado solo en PLN. Además, el método tuvo una alta velocidad de procesamiento de datos.

Estos trabajos muestran la diversidad y complejidad del análisis de sentimientos y discursos de odio en redes sociales, así como las diferentes técnicas y métodos que se han empleado para abordar este problema. Sin embargo, también existen otros campos y áreas donde estas técnicas pueden ser útiles y aplicables. Por ejemplo, Sri Mulyani et al. (2019) utilizaron el algoritmo Naïve Bayes para analizar los datos de los tweets sobre los programas de televisión en Indonesia, con el fin de obtener información sobre la valoración del sentimiento público. Los autores reportaron una precisión del 91.67% con su método, lo cual muestra el potencial de estas técnicas para otras industrias y sectores como: transporte y servicios públicos (Fahmi et al., 2023), salud (Alqarni & Rahman, 2023), educación (Dake & Gyimah, 2023), entre otras.

Sobre este trasfondo, en este estudio se propone el diseño y desarrollo de un sistema capaz de analizar y detectar contenido sexista en letras de canciones publicadas en medios digitales. Así, el estudio responde a la pregunta de investigación: ¿Es posible aplicar la minería de textos para la detección automática de sexismo en pistas musicales distribuidas en formato digital multimedia? Y de manera más específica se busca determinar lo siguiente: ¿cuáles son los algoritmos de aprendizaje automático que mejor precisión (*accuracy*) alcanzan para el análisis de sexismo en letras de canciones? En este contexto, el estudio que se propone ha experimentado con modelos basados en algoritmos como: Naïve Bayes, máquina de vectores de soporte, árbol de decisión, K-vecinos más cercanos y regresión logística (Pico-Valencia et al., 2021).

La detección de expresiones indecorosas de manera automática es importante, puesto que permite analizar el sentir del público con respecto a lo que observa y escucha. En el ámbito de reproductores de audio digital, la detección automática de frases sexistas podría ser útil para realizar un filtro de las canciones y así determinar si es adecuado o no para el público. El hecho de que los niños y jóvenes aprenden de lo que ven y escuchan, implica que no debe ser permisible que ellos tengan acceso a contenidos no acordes a su edad. Así, el sistema es pertinente debido a que puede ser empleado por padres de familia como mecanismo para evitar que sus hijos consuman audio digital con contenido vulgar y sexista.

La investigación se plantea como objetivo desarrollar un sistema que analice el nivel de sexismo en las letras de canciones del género urbano mediante el uso de técnicas de aprendizaje automático a fin de determinar si una canción es apta o no para determinado público (niños o adolescentes). Para el logro de este objetivo, se inició con la definición del corpus que describe acciones sexistas adaptado al vocablo ecuatoriano. Con base en este corpus, y considerando que se emplearon canciones publicadas en formato digital, se creó un mecanismo para transformar de manera automática contenido de audio digital a formato texto. Esto se efectuó para generar la entrada que requiere un clasificador automático. Estos datos fueron inyectados a un sistema inteligente que integra un modelo de detección automática de contenido sexista aplicando técnicas de minería de textos y aprendizaje automático. Finalmente, una vez entrenado el clasificador del sistema desarrollado, se aplicaron pruebas para validarlo. La evaluación del sistema propuesto se realizó en términos de la métrica exactitud. Para llevar a cabo la batería de pruebas se emplearon diez canciones del género urbano en español, disponibles en plataformas digitales de Internet. Dichas letras no se compartieron en ningún medio público, solo sirvieron de insumos para llevar a cabo la experimentación. Se citó la fuente respectiva en cada caso para no violar los derechos de autor.

El artículo está organizado en 3 secciones. La sección “Materiales y métodos” describe aspectos relacionados con el diseño del clasificador, así como los algoritmos de aprendizaje automático empleados en el experimento. La sección “Resultados y Discusión” presenta los hallazgos obtenidos en la fase de entrenamiento y evaluación del clasificador, que fue testeado con letras de canciones distintas a las usadas en la etapa de entrenamiento. Finalmente, la sección “Conclusiones” expone las principales conclusiones del estudio, así como los trabajos futuros que se originan de la investigación.

Materiales y Métodos

Aspectos metodológicos de la investigación

Esta investigación se caracterizó por ser de naturaleza mixta, abarcando tanto enfoques cualitativos como cuantitativos. La primera parte del estudio se concentró en realizar un análisis cualitativo de las técnicas basadas en datos y resultados previamente implementados por otros autores en el ámbito del aprendizaje automático, aplicado al análisis de sentimientos en redes sociales. La segunda etapa, de carácter cuantitativo, involucró la experimentación con diversos algoritmos que ayudaron a implementar la técnica de aprendizaje automático, con el propósito de evaluar la precisión y eficacia del clasificador entrenado. Esta evaluación cuantitativa se basó en las métricas obtenidas para entrenar y evaluar el clasificador entrenado.

En cuanto a las técnicas de procesamiento y análisis de datos, se utilizaron herramientas estadísticas para realizar un análisis más completo de los datos, incluyendo medidas de tendencia central y modelos probabilísticos implementados por los algoritmos de aprendizaje automático empleados en el desarrollo del sistema propuesto y en la experimentación; además, se utilizó la matriz de confusión como una técnica para evaluar el rendimiento del modelo de clasificación, permitiendo visualizar el número de predicciones correctas e incorrectas desglosadas por clase, lo que incluye verdaderos positivos, verdaderos negativos, falsos negativos y falsos positivos. Todas estas métricas fueron implementadas por la librería sklearn de Python.

Es importante señalar que esta investigación se enfocó en las técnicas de aprendizaje automático que son comúnmente empleadas para desarrollar sistemas de minería de texto destinados al análisis de sentimientos. A pesar de la amplia gama de algoritmos disponibles en el campo del aprendizaje automático, se optó por utilizar cinco algoritmos ampliamente reconocidos, basándose en la experiencia acumulada en investigaciones previas. Estos

algoritmos incluyen: Naïve Bayes, regresión logística, máquina de vectores de soporte, k-vecinos más cercanos, árbol de decisión y regresión logística (Pico-Valencia et al., 2021). Para llevar a cabo el análisis del sistema, se tomaron como referencia las canciones de los seis artistas de género urbano más populares durante el año 2020, según datos proporcionados por el portal digital accesible en la dirección (okdiario.com).

Algoritmos usados

Para llevar a cabo las tareas de aprendizaje por parte de las máquinas se emplean generalmente algoritmos que permiten predecir datos a través de regresiones, clasificadores y algoritmos de agrupamiento que se resumen en algoritmos de aprendizaje supervisado, no supervisado y por refuerzo (Cedeño-Moreno & Vargas, 2020)**.** El aprendizaje supervisado constituye el método que clasifica datos etiquetados con base a un patrón compartido. El propósito consiste en utilizar datos de entrenamiento, donde X representa las variables que anticipan una salida específica denominada Y. Estas variables pueden ser numéricas en el contexto de la regresión o descriptivas en situaciones de clasificación.

Los algoritmos de clasificación se usan cuando el resultado que se busca es una etiqueta discreta. Es decir, se tiene una clasificación binaria, solo se elige entre dos etiquetas, así mismo para la clasificación de múltiples etiquetas. A continuación, se describen los algoritmos más populares y usados en la actualidad para clasificar patrones. La descripción de cada algoritmo es muy específica y enfocada en describir el funcionamiento de cada algoritmo; no obstante, la forma en cómo se implementa cada uno de estos algoritmos en Python usando la librería sklearn se puede encontrar en el sitio oficial de la librería disponible en: https://n9.cl/50s6j.

Naïve Bayes

Es un clasificador lineal. Constituye un método de aprendizaje automático simple basado en probabilidades; las decisiones se toman en función de ausencia o presencia de determinadas características. Para ello, el algoritmo aplica el Teorema de Bayes con presunción de independencia entre dichas características. El clasificador Naïve Bayes combina el modelo de probabilidad derivado con una regla de decisión; es decir, se selecciona el valor, que tiene la mayor probabilidad. Este enfoque se conoce como el "Máximo a Posteriori" (Bijalwan et al., 2014).

Máquina de soporte vectorial

La Máquina de Vectores de Soporte (SVM por sus siglas en inglés) es un método de aprendizaje automático supervisado. SVM usa datos de entrenamiento para separar y construir un hiperplano de margen máximo que pueda usarse para la clasificación (Ramasamy et al., 2021). Las SVM se utilizan para mapear espacios de muestra en espacios de características con una dimensión alta (o incluso infinita) a través del mapeo no lineal, transformando los problemas que no son linealmente separables en el espacio muestral original y en problemas linealmente separables en el espacio de características (Xia et al., 2020). El concepto central detrás de las SVM es descubrir un hiperplano de margen máximo que logre una separación óptima entre las clases en el conjunto de datos. Los vectores de soporte corresponden a los puntos de datos que se hallan más cercanos al hiperplano; estos puntos desempeñan un papel crucial al definir de manera más precisa la línea de división al calcular los márgenes. Dichos puntos son esenciales para realizar la tarea de clasificación.

Vecino más cercano

El modelo k-vecinos más cercanos (k-NN por sus siglas en inglés) es un algoritmo sencillo que se usa tanto para problemas de regresión y clasificación. El algoritmo k-NN clasifica nuevas clases basados en medidas de similitud, asimismo, se lo utiliza para la

estimación de estadística y reconocimiento de patrones. El algoritmo se cataloga por mayoría de votos de sus vecinos, y el dato se asigna a la clase más común entre su vecino más cercano (k), el que es medido por una función de distancia. Para calcular la distancia para las variables continuas se aplica la distancia: euclidiana, manhattan o minkowski (Nugrahaeni & Mutijarsa, 2017).

Árbol de decisión

Es un modelo de clasificación alternativo que se asemeja a una estructura jerárquica de árbol, en la que cada nodo representa un atributo de prueba, las ramas reflejan los resultados de dichas pruebas y los nodos hoja denotan las clases (Apriliani et al., 2020). Además de su capacidad para manejar características de entrada y destino de tipo continuas, este modelo resulta útil en problemas con salidas categóricas. Su función principal radica en simplificar procesos de toma de decisiones complejos, permitiendo a los tomadores de decisiones comprender mejor la resolución de problemas mediante la identificación de características descriptivas que contienen información relevante (Nurfaizah et al., 2019).

Regresión logística

La regresión logística es un modelo estadístico empleado en el aprendizaje automático para describir las relaciones entre un conjunto de variables predictoras y una variable clase, con el propósito de estimar la probabilidad de que una instancia pertenezca a una clase específica. Si bien en su formulación original se usa como clasificador binario (predicción dicotómica), puede extenderse para abordar problemas de clasificación multiclase. Su denominación proviene de la función subyacente en la que se basa, conocida como función logística o sigmoide, y su tarea principal consiste en mapear las salidas de un modelo de regresión lineal en una probabilidad de pertenencia a una clase determinada (Wang et al., 2020).

Todos los algoritmos antes descritos permitieron entrenar clasificadores automáticos de frases sexistas a través de la librería sklearn de Python. Además, de manera complementaria, se empleó la librería NLTK (*Natural Language Toolkit*), que es una biblioteca ampliamente utilizada para PLN, proporcionando herramientas para trabajar con texto, siendo especialmente útil para tareas como tokenización, análisis sintáctico, clasificación de texto, entre otras.

Variables e indicadores sujetos a estudio

En el estudio se consideró una única variable: precisión del modelo. Dicha variable abarcó diversos indicadores, incluyendo la exactitud y la puntuación F1. Estas métricas se describen como sigue:

- Accuracy (Exactitud): Esta métrica mide el porcentaje de casos que el modelo ha clasificado correctamente.
- Precisión (Precisión): Se encarga de medir la calidad y exactitud del modelo de aprendizaje automático en sus predicciones.
- Recall (Recuperación): Evalúa la capacidad del modelo de reconocer y recuperar casos positivos.
- F1-Score (Puntuación F1): Es una métrica que combina las evaluaciones de precisión y recall en un único valor, ofreciendo una visión más completa del rendimiento del modelo.

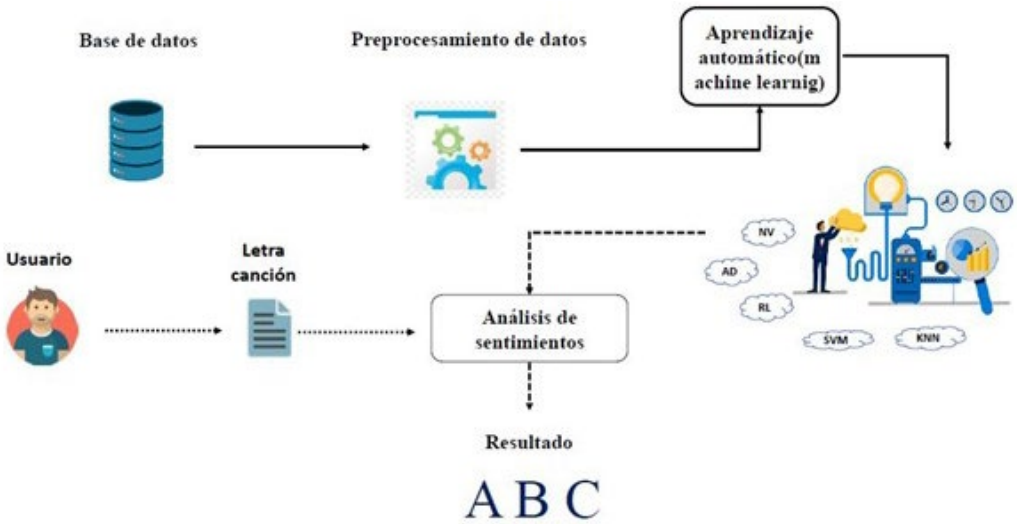
Diseño de los clasificadores para detección de frases sexistas

La arquitectura del sistema propuesto se fundamenta en la lógica de los sistemas de análisis de sentimientos. En la Figura 1 se puede observar que se tienen dos procesos. El primero corresponde al entrenamiento de los clasificadores; mientras que el segundo constituye el proceso de evaluación del modelo; este último es el que se lleva a cabo en entornos de producción.

Para realizar el primer proceso se inició con la creación del set de datos que sirvió de insumo para entrenar el modelo. Para crear el set de datos, se tomó como punto de referencia plataformas de música ampliamente reconocidas, como Deezer y Spotify, que gozan de reconocimiento a nivel mundial. El propósito principal fue identificar qué tipos de contenido musical son adecuados para audiencias específicas. Estas plataformas ofrecen listas de reproducción, conocidas como *playlists*, que ayudan a los usuarios a seleccionar sus contenidos favoritos.

Figura 1

Arquitectura del sistema propuesto



Se generó un conjunto de datos compuesto por 717 registros, los que se categorizaron en tres grupos: 1, 2 y 3. La categoría 1 corresponde con la designación "A" y representa contenido apto para todo público. La categoría 2, equivalente a "B," indica que el contenido requiere supervisión de un adulto, mientras que la categoría 3, equiparable a "C," se destina a contenido dirigido a mayores de edad o audiencias adultas. Esta clasificación se basó en el artículo 65 de la Ley Orgánica de Comunicación de Ecuador que establece que las personas naturales o jurídicas, propietarias o representantes legales de los medios de comunicación, responderán por los contenidos difundidos en los casos específicos mencionados. Algunos ejemplos de este tipo de contenidos consideran los siguientes: contenidos discriminatorios, que transgredan los derechos de las personas con VIH o personas LGBTI, que vulneren derechos de los niños, adultos mayores, personas discapacitadas, comunidades ancestrales, migrantes, entre otras (Ministerio Telecomunicaciones de Ecuador, 2019).

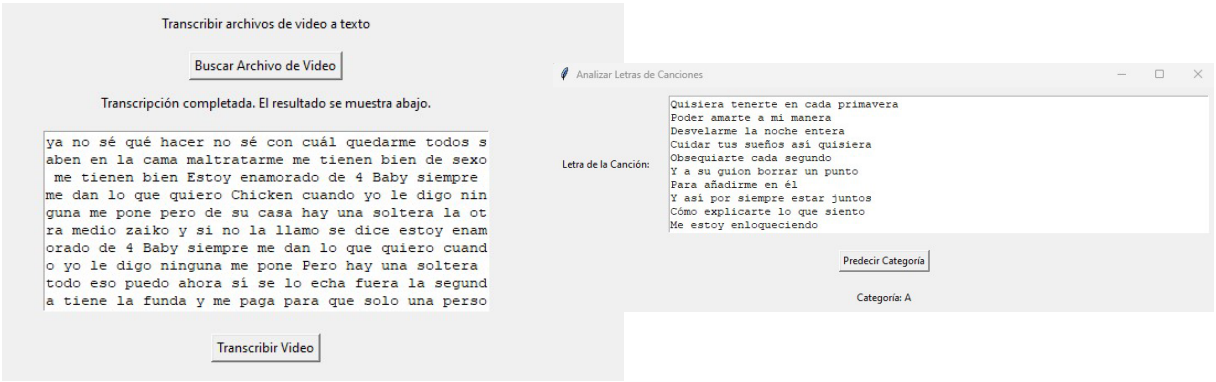
Posterior a la creación del set de datos se realizó su preprocesamiento a fin de que los datos estuvieran preparados para llevar a cabo el entrenamiento de los modelos de clasificación. El esquema del *set* de datos se organizó separando en dos columnas a la categoría de la letra y otra para las expresiones sexistas. Una vez que fueron preprocesados, a través de la técnica de ponderación TF-IDF para dar más importancia a palabras más relevantes y menos a las

comunes, estos se pasaron como argumentos para el entrenamiento de los clasificadores. Se creó un clasificador por cada uno de los siguientes algoritmos de aprendizaje supervisado: Naïve Bayes, máquina de vectores de soporte, k-vecinos más cercanos, árbol de decisión y regresión logística. Para el entrenamiento de los modelos de aprendizaje automático, se destinó el 75% de los datos para la fase de entrenamiento, reservando el 25% restante para las pruebas.

Por otro lado, en lo concerniente al proceso evaluación del modelo, luego de que los clasificadores fueron entrenados, se procedió a evaluar letras de canciones. La evaluación de las letras de canciones fue posible efectuarla especificando directamente la letra de la canción o en su defecto cargando el video digital de la canción. En este último caso, el proceso de conversión se realizó de forma automática (Figura 2 - izquierda) y, para ello, se empleó un script de Python. Dicho script utilizó las librerías *SpeechRecognition* y *MoviePy* para leer archivos de audio digital en formatos como MP4, convertir audio a texto y generar un archivo de texto con la transcripción. La librería probó su efectividad y precisión para trabajar con éxito en archivos cortos de calidad media. Complementariamente, en la Figura 2 derecha se muestra la interfaz gráfica desde la que el usuario puede llevar a cabo pruebas detallando la letra de la canción; proceso manual, válido en caso de que se disponga la letra de la canción en formato texto.

Figura 2

Interfaz de usuario para predecir categoría de letras de canciones



Resultados y Discusión

Métricas de los clasificadores en el proceso de entrenamiento

La Tabla 1 presenta los resultados obtenidos durante la evaluación de los modelos bajo una configuración específica, en la cual se utilizó el 75% de los datos para el proceso de entrenamiento y el 25% restante para las pruebas. Estos hallazgos revelan notables diferencias en el rendimiento de los modelos analizados. Para evaluar su desempeño, se emplearon diversas métricas de aprendizaje automático, entre las cuales se incluyen:

Tabla 1

Métricas del proceso de entrenamiento de los clasificadores

Algoritmos	Accuracy	Precision	Recall	F1
Naive Bayes (NB)	0.59	0.62	0.71	0.68
Regresión logística	0.75	0.77	0.80	0.74
Máquina de soporte vectorial	0.71	0.74	0.79	0.74
Árbol de decisión	0.70	0.68	0.76	0.73
k-Vecinos más cercanos	0.55	0.66	0.70	0.65

En este contexto, es relevante destacar que, según estos indicadores, dos modelos, el árbol de decisión y el modelo de k-vecinos más cercanos (k-NN), exhibieron un rendimiento menos satisfactorio en comparación con los demás; por el contrario, la regresión logística sobresalió al alcanzar la puntuación más alta en términos de precisión, lo que indica que es la opción más idónea para trabajar con el conjunto de datos específico en cuestión. Los resultados detallados de estas métricas se encuentran en la tabla 1, y estos hallazgos refuerzan la elección de la regresión logística como el modelo preferente para el procesamiento del conjunto de datos propuesto.

Métricas de los clasificadores en el proceso de evaluación

Con el propósito de desarrollar una evaluación más exhaustiva del sistema, se llevaron a cabo pruebas utilizando un conjunto de 10 canciones, cada una con contenido sexista y no sexista. Entre estos modelos se evidencia que, en el caso de la canción 3, los modelos Naïve Bayes, regresión logística y máquina de soporte vectorial, encasillan la canción como categoría 2, equivalente a la clasificación B. En contraste, el árbol de decisión la clasifica como 1, y k-vecinos más cercanos la cataloga como 3. Estas pruebas confirman que los dos últimos modelos presentan un margen de error más notable durante este proceso; además, los resultados destacan que solo cuando la letra de la canción pertenece a la clase 1, todos los modelos logran clasificarla de manera precisa.

La Tabla 2, que se presenta a continuación, resume los resultados de estas pruebas para cada una de las 10 canciones, mostrando las clasificaciones proporcionadas por los modelos evaluados. Estos hallazgos son fundamentales para comprender el rendimiento de los modelos en la clasificación de canciones con contenido sexista y no sexista.

Tabla 2

Métricas del proceso de evaluación de los clasificadores

Canción evaluada	Predicción Tipo	Naïve Bayes	Regresión Logística	Máquina de soporte vectorial	Árbol de decisión	k-Vecinos más cercanos
Four Babys (Maluma)	3	**3**	**3**	**3**	**3**	**3**
Yo te robaré (Ozuna)	2	**2**	**2**	**2**	**2**	**2**
Eres mía (Romeo Santos)	2	**2**	**2**	**2**	1	3
Primera cita (CNCO)	1	**1**	**1**	**1**	**1**	**1**
Desesperados (Raw)	3	**3**	**3**	**3**	1	**3**
Culpables (Anuel)	3	**3**	**3**	**3**	1	2
Hawái (Maluma)	2	**2**	3	3	1	**2**
Para enamorarte (CNCO)	1	**1**	**1**	**1**	**1**	**1**
Una vaina loca (Ozuna)	2	**2**	1	3	1	**2**
Bailando (Enrique Iglesias)	2	**2**	1	**2**	1	**2**

Conclusiones

La tecnología de análisis de sentimientos mediante el aprendizaje automático resulta altamente efectiva en la detección de sexismo, logrando una destacada precisión del 77% a partir del modelo de aprendizaje propuesto implementado a través del algoritmo de regresión logística. El modelo desarrollado puede mejorarse ampliando el corpus usado para su entrenamiento. Por el hecho de que estos modelos son dependientes de los datos se evidenció

que con la incorporación de más datos se puede mejorar la precisión de los algoritmos. En este sentido, se recomienda extender el set de datos de entrenamiento con jergas iberoamericanas.

Es importante señalar que la mayoría de los algoritmos de aprendizaje automático, usados en la etapa de experimentación, demostraron una alta precisión, con la excepción de los algoritmos de árbol de decisión y vecinos más cercanos. Los dos algoritmos que lograron mejor precisión en la detección de sexismo fueron la regresión logística con el 77 % y la máquina de vectores de soporte con el 74 %. La capacidad actual del modelo entrenado mediante el uso de la regresión logística es aceptable y, por tanto, este puede ser usado para que los padres puedan evaluar si el contenido lírico de una canción es apropiado o no para sus hijos; además, sienta las bases para desarrollar modelos más precisos a partir de redes neuronales de manera que sean integrados como mecanismos de supervisión en plataformas como YouTube o Spotify.

Un aporte importante del estudio es que contribuye con el cumplimiento de la Ley Orgánica de Comunicación de Ecuador que en el artículo 76 menciona que los sistemas de audio y video deben estar previamente calificados por el Consejo de Regulación y Desarrollo de la Información y Comunicación, considerando la calidad de sus contenidos y programación, siempre que satisfagan las condiciones técnicas que establezca la autoridad de telecomunicaciones (Ministerio Telecomunicaciones de Ecuador, 2019). En este sentido, el modelo entrenado aprendió a categorizar letras de canciones en las tres categorías contempladas en este estudio: "A", contenido apto para audiencias de todas las edades; "B", contenido que requiere supervisión de adultos; y "C", apto para adultos.

Como perspectiva para investigaciones futuras, se considera la extensión de la propuesta hacia la capacidad de realizar predicciones en tiempo real. Esto permitirá que las predicciones se ajusten de manera dinámica a medida que avanza la reproducción de pistas de audio o vídeos digitales. Esta ampliación tendría importantes aplicaciones prácticas en la verificación de contenidos publicados en redes sociales, medios que son considerablemente usados por los niños y adolescentes en la actualidad. También la propuesta puede extenderse para que sea multilingüe de manera que contemple el análisis de frases sexistas en español e inglés; así como en otros posibles géneros musicales.

Reconocimientos

Esta investigación es fruto de uno de los trabajos de titulación de la Escuela de Sistemas de la Pontificia Universidad Católica del Ecuador, sede Esmeraldas, enmarcada en la línea de investigación Automatismos y sistemas inteligentes.

Referencias

Alqarni, A., & Rahman, A. (2023). Arabic Tweets-Based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach. *Big Data and Cognitive Computing*, *7*(1), 1–29. https://doi.org/10.3390/bdcc7010016

Apriliani, D., Abidin, T., Sutanta, E., Hamzah, A., & Somantri, O. (2020). Sentiment analysis for assessment of hotel services review using feature selection approach based-on decision tree. *International Journal of Advanced Computer Science and Applications*, *11*(4), 240–245. https://doi.org/10.14569/IJACSA.2020.0110432

Arce-García, S., & Menéndez-Mendéndez, M.-I. (2023). Inflamando el debate público: metodología para determinar origen y características de discursos de odio sobre diversidad sexual y de género en Twitter and gender diversity on Twitter. *Profesional de La Información*, *3*(1), 1–19. https://doi.org/10.3145/epi.2023.ene.06

Back, B. H., & Ha, I. K. (2019). Comparison of sentiment analysis from large twitter datasets by naive bayes and natural language processing methods. *J. Inf. Commun. Converg. Eng.*, *17*(4), 239–245. https://doi.org/10.21541/apjes.939338

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), 61–70. https://doi.org/10.14257/ijdta.2014.7.1.06

Castañeda Muñoz, J. (2019). *Análisis, clasificación y predicción del vocabulario de cibercrimen en Internet usando modelos predictivos de Machine Learning* [Tesis de Maestría, Universidad Cuahtémoc]. https://uconline.mx/comunidadead/application/views/repositoriodetesis/TesisfinalJoseAlexanderCastanedaMunoz.pdf

Cedeño-Moreno, D., & Vargas, M. (2020). Aprendizaje automático aplicado al análisis de sentimientos. *I+D Tecnológico*, *16*(2), 59–66. https://doi.org/10.33412/idt.v16.2.2833

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies*, *28*(4), 4629–4647. https://doi.org/10.1007/s10639-022-11349-1

Dhrodia, A. (2017). *Social media and the silencing effect: why misogyny online is a human rights issue*. https://www.newstatesman.com/culture/social-media/2017/11/social-media-and-silencing-effect-why-misogyny-online-human-rights-issue

Fahmi, M., Yuningsih, Y., & Puspita, A. (2023). Sentiment Analysis Of Online Gojek Transportation Services On Twitter Using The Naïve Bayes Method. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *8*(2), 84–90. https://doi.org/10.33480/jitk.v8i2.4004

Ghosh, S., Roy, S., & Bandyopadhyay, S. K. (2012). A tutorial review on Text Mining Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(4), 223–233.

Jiang, L., & Suzuki, Y. (2019). Detecting hate speech from tweets for sentiment analysis. *2019 6th International Conference on Systems and Informatics, ICSAI 2019*, *Icsai*, 671–676. https://doi.org/10.1109/ICSAI48974.2019.9010578

Lepe, M. (2021). *Modelos híbridos basados en Lexicones y Machine Learning para la detección de agresividad sobre textos en idioma Español*. https://www.mcc.ubiobio.cl/web/docs/tesis/manuel_lepe-tesis(manuellepe).pdf

Mesiti, A. M., & Yeo, H. L. (2023). Social Media: The Good, the Bad, and the Ugly. *Clinics in Colon and Rectal Surgery*, *36*(5), 347–352. https://doi.org/10.1055/s-0043-1763281

Ministerio Telecomunicaciones de Ecuador. (2019). *Ley Orgánica de Comunicaciones*. https://www.telecomunicaciones.gob.ec/wp-content/uploads/2020/01/Ley-Organica-de-Comunicación.pdf

Nugrahaeni, R. A., & Mutijarsa, K. (2017). Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification. *Proceedings - 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, 163–168. https://doi.org/10.1109/ISEMANTIC.2016.7873831

Nurfaizah, Hariguna, T., & Romadon, Y. I. (2019). The accuracy comparison of vector support machine and decision tree methods in sentiment analysis. *Journal of Physics: Conference Series*, *1367*(1). https://doi.org/10.1088/1742-6596/1367/1/012025

OED. (1866). *Sexism*. https://www.oed.com/search/dictionary/?scope=Entries&q=sexism

Penagos Rojas, Y. (2012). Lenguajes del poder. La música reggaetón y su influencia en el estilo de vida de los estudiantes. *Plumilla Educativa*, *10*(2), 290–305. https://dialnet.unirioja.es/servlet/articulo?codigo=4323457

Pico-Valencia, P., Vinueza-Celi, O., & Holgado-Terriza, J. A. (2021). Bringing Machine Learning Predictive Models Based on Machine Learning Closer to Non-technical Users. *Advances in Intelligent Systems and Computing*, *1273 AISC*, 3–15. https://doi.org/10.1007/978-3-030-59194-6_1

Piñón Lora, M., & Pulido Moreno, A. (2020). La imagen de la mujer en el reggaetón: un análisis crítico del discurso. *Revista Iberoamericana de Comunicación*, *38*, 45–77. https://ric.ibero.mx/index.php/ric/article/view/67/53

RAE. (2023). *Sexismo*. http://dle.rae.es/srv/search?m=30&w=sexismo

Ramasamy, L. K., Kadry, S., & Lim, S. (2021). Selection of optimal hyper-parameter values of support vector machine for sentiment analysis tasks using nature-inspired optimization methods. *Bulletin of Electrical Engineering and Informatics*, *10*(1), 290–298. https://doi.org/10.11591/eei.v10i1.2098

Rasel, R. I., Sultana, N., Akhter, S., & Meesad, P. (2018). Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach. *ACM International Conference Proceeding Series*, 37–41. https://doi.org/10.1145/3278293.3278303

Sri Mulyani, E. D., Rohpandi, D., & Rahman, F. A. (2019). Analysis of Twitter Sentiment Using the Classification of Naive Bayes Method about Television in Indonesia. *2019 1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, *1*(August), 89–93. https://doi.org/10.1109/ICORIS.2019.8874896

Wang, P., Yan, Y., Si, Y., Zhu, G., Zhan, X., Wang, J., & Pan, R. (2020). Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text. *IEEE Access*, *8*, 97370–97382. https://doi.org/10.1109/ACCESS.2020.2995905

Xia, H., Yang, Y., Pan, X., Zhang, Z., & An, W. (2020). Sentiment analysis for online reviews using conditional random fields and support vector machines. *Electronic Commerce Research*, *20*(2), 343–360. https://doi.org/10.1007/s10660-019-09354-7

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10843 LNCS*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93417-4_48