ISSN 2007-9737

Identificación automática de contenido misógino en redes sociales: Un enfoque basado en transferencia de conocimiento proveniente de canciones

Ricardo Calderón-Suarez[1,3], Rosa María Ortega-Mendoza[1], Marco Antonio Márquez-Vera[2], Félix Agustín Castro-Espinoza[*,1]

[1] Universidad Autónoma del Estado de Hidalgo, Hidalgo, México

[2] Universidad Politécnica de Pachuca, Hidalgo, México

[3] Universidad Politécnica de Tulancingo, Hidalgo, México

{ricardo_calderon, rosa.ortega, fcastro}@uaeh.edu.mx marquez@upp.edu.mx,

Resumen. Este artículo de investigación presenta un resumen de la tesis "Detección Automática de Contenido Misógino en Redes Sociales mediante Transferencia de Conocimiento proveniente de Canciones", donde la idea principal es aprovechar el conocimiento existente en algunas canciones para transferir patrones lingüísticos que ayuden a identificar manifestaciones de misoginia en las redes sociales. En particular, se analizaron varias técnicas de transferencia de aprendizaje. Además, se presenta una metodología para construir, automáticamente, una colección de canciones y otra de frases, ambas con instancias etiquetadas de acuerdo con la presencia o ausencia de contenido misógino. La mayor contribución de esta investigación es un método de aumentación de datos que incrementa la capacidad de generalización de los modelos de detección de misoginia mediante la transferencia de la riqueza semántica contenida en las letras de las canciones. El enfoque propuesto fue evaluado en colecciones de referencia que contienen textos en español e Inglés, obteniendo resultados alentadores. En comparación con enfoques robustos del estado del arte, el enfoque propuesto obtuvo resultados competitivos en el idioma Inglés y ganancias importantes en el idioma Español. Esta investigación confirmó la existencia de conocimiento lingüístico valioso en las canciones, el cual puede ser transferido para detectar contenido misógino en redes sociales.

Palabras clave. Transferencia de aprendizaje, aumentación de datos, detección de misoginia, redes sociales.

Automatic Identification of Misogynistic Content on Social Networks: An Approach based on Knowledge Transfer from Songs

Abstract. This research paper presents a summary of the thesis "Automatic Detection of Misogynistic Content in Social Networks through Knowledge Transfer from Songs", where the main idea is to leverage the existing knowledge of some songs to transfer linguistic patterns that help to identify manifestations of misogyny in social media. In particular, several learning transfer techniques were analyzed. In addition, a methodology

is presented to build, automatically, a collection of songs and another of phrases, both with instances labeled according to the presence or absence of misogynistic content. The major contribution of this research is a data augmentation method that increases the generalization capability of the misogyny detection models by transferring the semantic richness contained in song lyrics. The proposed approach was evaluated in benchmark collections containing texts in Spanish and English, obtaining encouraging results. Compared to robust state-of-the-art approaches, the proposed approach obtained competitive results in English and significant gains in Spanish. This research confirmed the existence of valuable linguistic knowledge in songs, which can be transferred to detect misogynistic content in social media.

Keywords. Transfer learning, data augmentation, mysogyny detection, social media.

1. Introducción

La misoginia ha lastimado seriamente el bienestar de la sociedad y en casos severos, ha conducido a feminicidios [26]. Este concepto abarca ideas culturales que sugieren la inferioridad de las mujeres y se manifiesta a través de diversas formas como el menosprecio, la discriminación de género, acoso sexual, objetivación sexual, violencia verbal y física contra las mujeres [12]. Este comportamiento ha estado presente en la sociedad y ha evolucionado con el tiempo de acuerdo con el contexto social, cultural y religioso de los países o regiones.

Hoy en día, las manifestaciones de misoginia se observan en distintos niveles y en diversos medios de comunicación, tales como la música y las plataformas sociales (e.g., Facebook y Reddit) [21, 26, 33]. La misoginia se ha estudiado desde diversas áreas como la sociología y la psicología [18, 12]. En particular, se ha establecido una relación entre la misoginia y el lenguaje [32].

En este contexto, en el año 2016 se presentó un estudio sobre el uso del lenguaje misógino en redes sociales, encontrando hallazgos interesantes como la escritura frecuente del título de algunas canciones populares [22]. Más tarde, se publicó un trabajo pionero sobre la detección y clasificación automática de misoginia en tweets [3].

A la fecha, se han realizado varios esfuerzos para identificar automáticamente contenido ofensivo contra las mujeres en publicaciones provenientes de diversas plataformas sociales [16, 18, 24]. Recientemente, se han creado foros internacionales para evaluar métodos automáticos que abordan la tarea de identificación automática de misoginia en plataformas sociales, la cual es conocida como AMI (por sus siglas en Inglés, Automatic Misogyny Identification).

En el año 2018, dentro de los foros Ibereval [18] y Evalita [17], se lanzó una tarea compartida para identificar y clasificar mensajes misóginos escritos en Inglés, Español e Italiano de la red social Twitter[1]. En general, los enfoques participantes exploraron representaciones basadas en n-gramas y embeddings así como diversas características lingüísticas, tales como léxicas y estilísticas.

Actualmente, la tarea se ha extendido hacia la detección de contenido misógino en escenarios multimodales, implicando el procesamiento de texto e imágenes. En este contexto, en el año 2022, dentro del foro SemEval, se estableció la tarea denominada MAMI (por sus siglas en Inglés, Multimedia Automatic Mysogyny Identification) [16], la cual fue dirigida hacia la detección de misoginia en memes.

En esta competencia, el uso de modelos pre-entrenados para tratar texto e imágenes fue un factor común. La tarea AMI ha sido comúnmente abordada desde un marco de clasificación supervisada de textos.

Por lo tanto, el desempeño de los clasificadores depende en gran medida del tamaño, así como de la calidad de los conjuntos de datos de entrenamiento. Sin embargo, actualmente esta tarea se enfrenta a la escasez de colecciones de datos de entrenamiento etiquetados.

Además, resulta complicado encontrar conjuntos que contengan todo el vocabulario que exprese actitudes misóginas, tanto de forma implícita como explícita. Por ejemplo, la detección del lenguaje abusivo explícito se enfrenta a diferentes desafíos, como la identificación de vocabulario informal o jerga, así como la diversidad de significados de algunos

[1] Actualmente, esta plataforma ha cambiado su nombre a X.

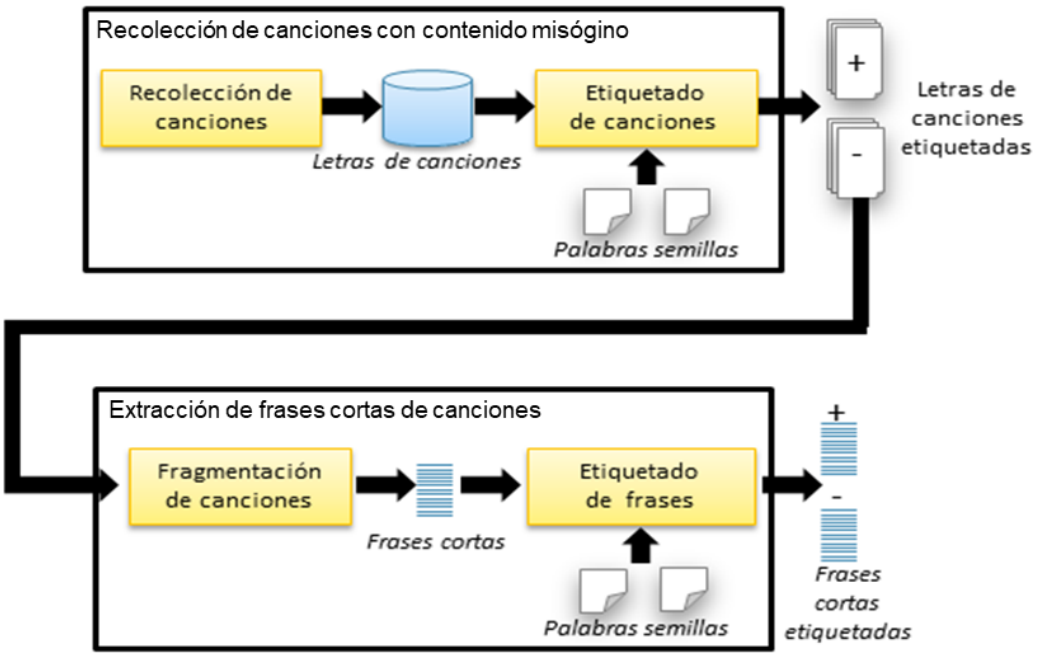


Fig. 1. Metodología propuesta para la construcción de colecciones de canciones y frases etiquetadas según la presencia de contenido misógino

términos relevantes (e.g., ofensas y blasfemias) dependiendo del contexto donde son usados. Por otro lado, identificar el lenguaje abusivo implícito también presenta dificultades, ya que las agresiones pueden estar ocultas o disfrazadas a través de chistes, bromas o comentarios sarcásticos [30, 34]. Tratando de enfrentar los desafíos mencionados, en este trabajo de investigación se propone enriquecer la capacidad de generalización de los modelos de detección de misoginia.

En particular, se desarrolló un enfoque para identificar manifestaciones de misoginia en redes sociales mediante la transferencia de conocimiento proveniente de otro dominio, específicamente de las letras de canciones.

El objetivo es agregar nueva información para enriquecer la diversidad de los patrones lingüísticos encontrados durante el entrenamiento. Las siguientes preguntas de investigación motivaron el trabajo: ¿Las canciones contienen patrones lingüísticos que pueden ser explotados por modelos de detección de misoginia en el ámbito de las redes sociales?, ¿Las frases de canciones pueden ser aprovechadas para aumentar los datos de entrenamiento? y ¿El enfoque propuesto puede ser adaptado para detectar misoginia en contextos multimodales?

Con el propósito de investigar las respuestas, se diseñó un enfoque que automáticamente detecta contenido misógino en redes sociales, empleando transferencia de conocimiento derivado de letras de canciones. El resto del manuscrito resume la investigación de la tesis [9] y las publicaciones derivadas [10].

2. Construcción automática de los conjuntos de datos: etiquetando canciones

La música se ha considerado como un medio de comunicación masiva en el que se transmiten ideas, sentimientos y emociones [7, 14, 19]. En este sentido, se ha encontrado que la música está vinculada al contexto donde se produce, lo que motivado diferentes investigaciones.

ISSN 2007-9737

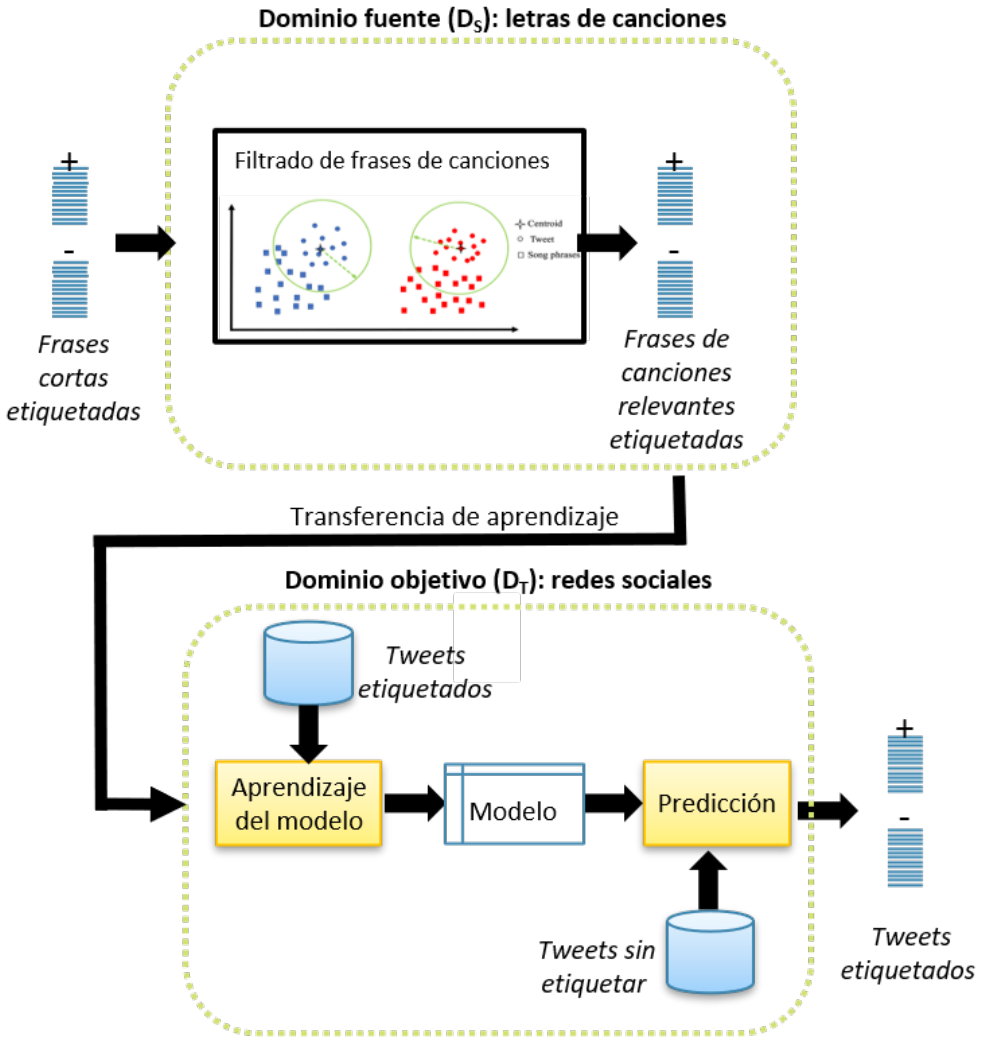


Fig. 2. Esquema del enfoque propuesto para aumentación de datos

Por ejemplo, algunos hallazgos indican que existen letras de canciones populares que expresan una representación desfavorable de las mujeres [4, 6], mostrando algunos fenómenos arraigados en la sociedad, tales como: objetivación sexual de las mujeres [33], inferioridad femenina e incluso violencia contra las mujeres [1, 8].

Por lo tanto, el contenido de las canciones y el uso del lenguaje en tal dominio pueden constituir una fuente valiosa de conocimiento para detectar, de manera automática, comportamiento agresivo contra las mujeres.

Para explorar y aprovechar esta base de conocimientos, en esta investigación se diseñó una metodología que recopila y etiqueta de manera automática dos conjuntos de datos: uno compuesto por canciones completas y otro por las frases que contienen expresiones de misoginia extraídas de dichas canciones.

El proceso completo para generar las colecciones mencionadas en el párrafo anterior se muestra en la Figura 1. Como se observa, tanto las canciones como las frases son etiquetadas mediante las categorías: misógina y no misógina.

ISSN 2007-9737

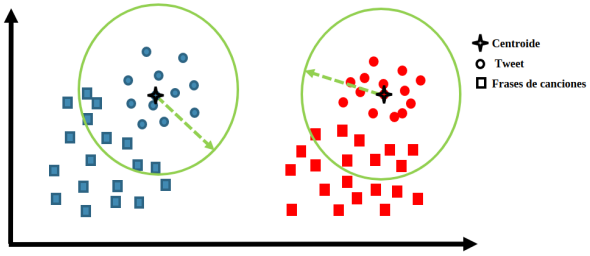


Fig. 3. Ilustración del mecanismo de filtrado propuesto. La calidad de las instancias está asociada con la distancia hacia el correspondiente centroide. Los tweets positivos y negativos se encuentran representados con círculos rojos y azules, respectivamente. Las frases de canciones están ilustradas mediante pequeños cuadrados

Tabla 1. Distribución de las colecciones generadas. Se muestran las estadísticas del conjunto de canciones (Ca) y Frases (Fr). Las etiquetas fueron asignadas automáticamente y corresponden a la categoría misógina (M) o no misógina (N)

Idioma	Conjunto	M	N	Total
Español	Ca	4228	4228	8456
	Fr	1411	1411	2822
Inglés	Ca	11086	11086	22172
	Fr	2120	2120	4240

En las siguientes secciones se describen las etapas de la metodología propuesta.

2.1. Recolección de canciones con contenido misógino

En esta etapa, se recopilan y etiquetan automáticamente canciones. Durante la recolección, se incluyeron canciones que han sido señaladas como misóginas por activistas en diversos foros web. También se añadieron otras seleccionadas aleatoriamente.

Además, con el fin de asegurar un amplio vocabulario de términos, se recolectaron composiciones de diversos autores y estilos musicales, las cuales fueron obtenidas de distintas plataformas en línea[2].

[2]Por ejemplo: www.lyrics.com/ y www.letras.com/

Un paso importante dentro de la metodología es etiquetar automáticamente las letras de canciones (i.e., asignar una etiqueta a cada instancia). Para ello, se diseñó un proceso automático que considera la presencia de palabras clave, a las cuales se les denomina semillas.

Específicamente, se emplearon dos tipos de semillas: palabras misóginas y palabras relacionadas con el término mujer. Las palabras misóginas son aquellas asociadas con la manifestación de abuso verbal contra la mujer.

Para esta investigación se emplearon dos léxicos de términos vinculados con misoginia, los cuales fueron presentados en [15, 27], para el idioma Inglés y Español, respectivamente. Por otro lado, las semillas relacionadas con el término mujer se utilizaron para garantizar que el contenido de las letras haga referencia a las mujeres.

En este contexto, se consideró una lista de palabras clave comúnmente relacionadas[3], tales como niña, novia y esposa. Considerando estas semillas, los criterios diseñados para definir las etiquetas de cada canción son los siguientes:

Misógina. Se asigna esta etiqueta cuando una canción contiene palabras de ambos tipos de semilla: una relacionada con mujer y al menos dos palabras vinculadas con misoginia.

No misógina. Se asigna esta etiqueta a las canciones que no contienen palabras misóginas.

2.2. Extracción de frases cortas de canciones

La segunda etapa de la metodología propuesta está orientada a extraer frases cortas de canciones y etiquetarlas con las categorías misógina o no misógina. El proceso inicial consiste en fragmentar las letras de las canciones en segmentos, cada uno de los cuales tiene una extensión máxima de 280 caracteres.

Cabe señalar que se eligió un tamaño de longitud similar a aquella de las publicaciones en la plataforma Twitter, ya que el método se enfoca en detectar misoginia en tweets.

[3]Fueron obtenidas consultando los siguientes portales web: relatedwords.org para la configuración en el idioma Inglés y www.ideasafines.com.ar/do-buscar.php para el idioma Español.

Tabla 2. Algunas estadísticas de los conjuntos de frases filtradas utilizando las técnicas propuestas basadas en similitud Coseno (F-Coseno) y el algoritmo de Roccio (F-Roccio)

Datos	Conjunto	Misógino	No Misógino	Total
Español	Frases	1411	1411	2822
	F-Coseno	282	282	564
	F-Roccio	290	1411	1701
Inglés	Frases	1783	1783	3566
	F-Coseno	357	357	714
	F-Roccio	1085	1776	2861

Los criterios de etiquetado de las frases siguen un procedimiento similar al etiquetado de las canciones completas. La etiqueta misógina (clase positiva) se otorga a aquellas frases provenientes de canciones etiquetadas con esta categoría y que, además, contienen dos palabras semillas vinculadas con la misoginia y una relacionada con el término mujer. En contraste, la etiqueta No misógina (clase negativa) se otorga a frases cortas elegidas aleatoriamente del conjunto de canciones no misóginas.

2.3. Resultados de la construcción de las colecciones

Las estadísticas de las colecciones resultantes del proceso de etiquetado de canciones y la extracción de frases se muestran en la Tabla 1. Como se observa, se crearon colecciones de acuerdo con el idioma en el que fueron escritas las canciones (Español o Inglés).

Como parte del análisis de las colecciones construidas, se realizó una exploración de su vocabulario. En general, se observó que los términos más frecuentes incluyen, además de los términos semilla, palabras despectivas u ofensivas contra las mujeres.

También, se observaron referencias a diversas partes del cuerpo, la cuales son comúnmente relacionadas con el concepto de cosificación sexual. En general, el análisis indica que los fragmentos de canciones que contienen palabras semilla muestran manifestaciones de misoginia.

Este conocimiento lingüístico puede ser relevante para diversas tareas basadas en clasificación automática, por ejemplo, la detección de misoginia.

3. Un nuevo enfoque de aumentación de datos usando frases de canciones

La tarea AMI suele abordarse bajo un esquema de clasificación de textos. En este enfoque, la calidad de los clasificadores se relaciona con su capacidad para generalizar, la cual, a su vez, depende de la cantidad de datos de entrenamiento. Sin embargo, esta tarea se ha enfrentado a la poca disponibilidad de datos etiquetados para entrenar los modelos computacionales.

Actualmente, una de las soluciones al problema de escasez de datos de entrenamiento etiquetados contempla el uso de técnicas de aumentación de datos [23]. En este trabajo de investigación se propone un enfoque de aumentación de datos que aprovecha el conocimiento y patrones lingüísticos provenientes de las canciones. La Figura 2 muestra una visión general del método propuesto.

Su objetivo es utilizar frases de alta calidad de las canciones para aumentar los conjuntos de entrenamiento relacionados con la tarea. La idea clave es incrementar la capacidad de aprendizaje de los modelos diversificando las instancias de entrenamiento con ejemplos de expresiones socioculturales contenidas en la música.

3.1. Transfiriendo conocimiento proveniente de las canciones

El enfoque propuesto se sitúa en el dominio de los enfoques de transferencia de aprendizaje, ya que aprovecha el conocimiento existente en las canciones para utilizarlo en una tarea fuera del dominio (out-domain). En concordancia con las notaciones en [2], el concepto de transferencia de aprendizaje se define enseguida. Sean D_S datos del dominio fuente, D_T datos del dominio objetivo o destino, T_S la tarea de aprendizaje del dominio fuente y T_T representa la tarea de aprendizaje en el dominio objetivo.

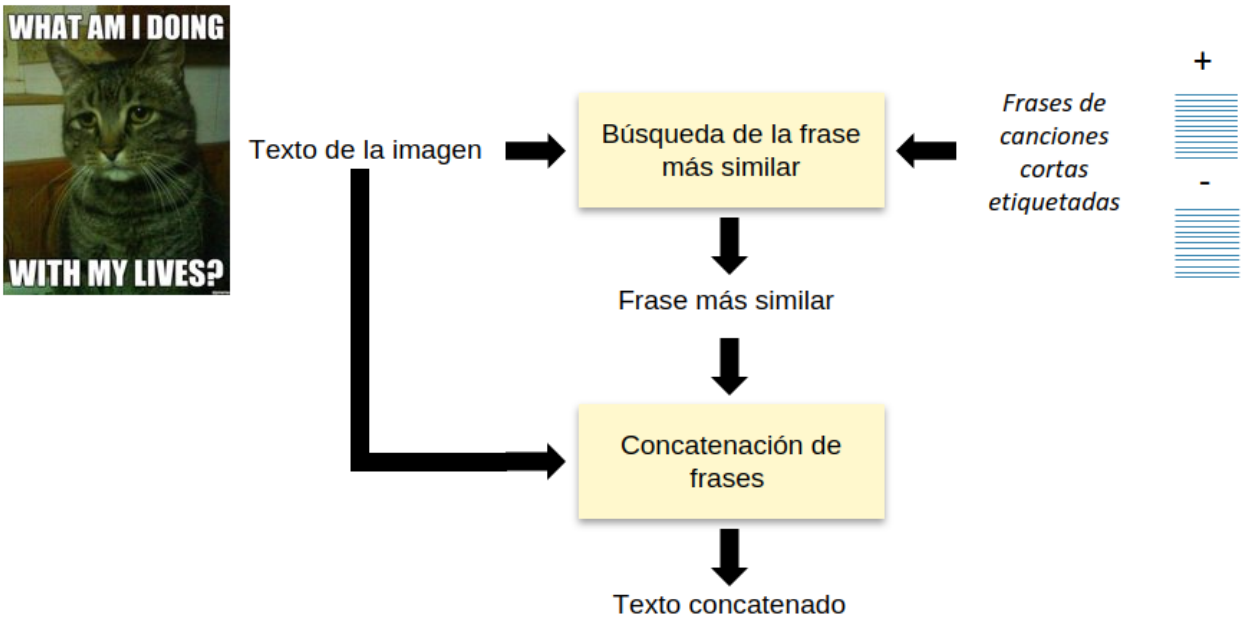


Fig. 4. Representación de la técnica propuesta para expandir el texto contenido en los memes. La imagen fue extraída del conjunto de entrenamiento del concurso SemEval 2022 (tarea MAMI)

La meta del aprendizaje por transferencia es emplear el conocimiento del dominio fuente y su tarea asociada en el proceso de aprendizaje para la tarea del dominio objetivo, donde $D_S \neq D_T$ o $T_s \neq T_T$.

Para este trabajo de investigación, las letras de canciones y las redes sociales son consideradas los dominios D_S y D_T, respectivamente, mientras $T_S = T_T$ (i.e., detección de misoginia).

El método propuesto aumenta los datos de entrenamiento del D_T siguiendo un enfoque de dominio cruzado. El objetivo es agregar únicamente frases de calidad que aporten en la mejora del desempeño del clasificador. Para lograr esto, se diseñó un mecanismo de filtrado, el cual es descrito en la siguiente sección.

3.2. Mecanismo de filtrado

La variedad de temas presentes en las canciones puede generar oraciones que no contribuyen al proceso de generalización, añadiendo ruido y afectando el rendimiento del clasificador.

Para solventar este problema, se propone un mecanismo que evalúa la calidad de las frases y selecciona solamente aquellas más pertinentes para la tarea (es decir, las de mayor calidad). En particular, se diseñó un filtro basado en la similitud de las frases de canciones con el conjunto de entrenamiento formado por los tweets.

La Figura 3 esquematiza el filtro propuesto. Su propósito consiste en seleccionar únicamente las instancias del dominio de fuente que estén más próximas a los centroides de cada clase existente en el dominio destino.

Para cuantificar la proximidad existente entre los centroides de los tweets y las frases de las canciones, se empleó la similitud Coseno. En este sentido, se sugiere retener únicamente un porcentaje (θ) de las frases con mayor similitud.

Cabe señalar que, también se empleó el clasificador Roccio[4] como una estrategia alternativa al uso de la similitud Coseno.

[4] scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestCentroid.html

Tabla 3. Desempeño de una BoW aplicando diversas configuraciones de adaptación de dominio. Los resultados se reportan en términos de exactitud (Exac) y F_1. El método de referencia corresponde a un clasificador entrenado únicamente con los tweets (Tw). La información de las letras se consideró en el entrenamiento a partir de las dos colecciones construidas: letras de canciones completas (LetraC) y frases de canciones (Frases)

Datos		MVS		XG		RL	
		Exac	**F1**	**Exac**	**F1**	**Exac**	**F1**
Iber-Es	Tw	**0.819**	**0.819**	0.781	0.780	0.813	0.813
	LetraC	0.659	0.646	0.656	0.648	0.651	0.629
	Frases	0.665	0.656	0.657	0.649	0.649	0.641
Iber-In	Tw	**0.806**	0.793	0.798	0.772	0.762	0.723
	LetraC	0.665	0.591	0.675	0.574	0.678	0.601
	Frases	0.686	0.633	0.669	0.574	0.697	0.654
Eval-In	Tw	0.597	0.597	0.567	0.562	0.606	0.602
	LetraC	0.638	0.634	0.642	0.627	**0.644**	0.636
	Frases	0.615	0.614	0.641	0.633	0.623	0.622

3.3. Colecciones filtradas: Resultados del proceso de filtrado

Los conjuntos conformados por las instancias filtradas se muestran en la Tabla 2. La colección generada usando el filtro basado en similitud Coseno fue intencionalmente balanceada con respecto al número de instancias de la clase positiva. Los conjuntos de frases filtradas serán utilizadas para aumentar los datos del entrenamiento.

4. Adaptación para detectar misoginia en ambientes multimodales

Como parte de la investigación, el enfoque propuesto fue adaptado para detectar misoginia en ambientes multimodales. La adaptación propuesta fue evaluada con datos la tarea MAMI del foro Semeval 2022 [16], cuyo objetivo fue detectar memes con contenido misógino.

La idea principal de la adaptación del enfoque es transferir el conocimiento de las canciones hacia la clasificación de memes, los cuales pueden incluir texto. Considerando que el texto de los memes presenta, comúnmente, una longitud corta, se propone expandirlo añadiendo frases muy similares provenientes de las canciones.

De esta manera, se agrega nueva información que puede enriquecer los patrones lingüísticos discriminativos para la tarea. El procedimiento de expansión se encuentra representado en la Figura 4.

Como se observa, antes de ingresar los memes al clasificador, pasan por el proceso de expansión. Posteriormente, las instancias son clasificadas usando una arquitectura multimodal. Particularmente, para los experimentos, se utilizó el modelo presentado en [31].

5. Configuración experimental

En las siguientes secciones se muestra el marco de trabajo experimental: conjuntos de datos usados, representaciones textuales y clasificadores.

5.1. Conjuntos de datos

Las ideas de este trabajo de investigación fueron evaluadas usando los conjuntos de datos provenientes de los foros: Ibereval [18] y Evalita [17]. Particularmente, se realizaron experimentos con los datos en Español e Inglés del primer conjunto, mientras que de Evalita se utilizaron los datos en Español.

Tabla 4. Comparación de desempeño (exactitud) de diferentes modelos de clasificación empleando los embedings generales (Gen) y especializados (Esp). Los vectores de palabras fueron evaluados a través de dos representaciones de texto: vector promedio (AWE) y como capa de entrada de un modelo GRU

Conjunto	Tipo	AWE			GRU
		MVS	XG	RL	Prom±DE
Iber-Es	Gen	0.776	0.781	0.762	0.779 ± 0.012
	Esp	**0.788**	0.771	0.777	**0.792±0.018**
Iber-In	Gen	0.751	0.731	0.729	0.729 ± 0.029
	Esp	**0.792**	0.758	0.791	0.742 ± 0.022
Eval-In	Gen	0.641	0.625	**0.665**	0.576 ± 0.018
	Esp	0.615	0.617	0.618	0.588 ± 0.026

En las siguientes secciones se hará referencia a ellos a través de la siguiente notación: Iber-Es, Iber-In y Eval-In, respectivamente. Para el escenario multimodal se utilizó el conjunto de datos proveniente de la tarea Mami en el foro SemEval 2022 [16].

5.2. Representaciones textuales

Pre-procesamiento. El texto de los tweets fue procesado como sigue: conversión a minúsculas, eliminación de palabras vacías, así como caracteres especiales, emojis, URL's y las menciones a usuarios. Además, el texto fue separado en unigramas de palabras y para crear las representaciones textuales se usaron los 10,000 términos más frecuentes.

Bolsa de palabras BoW. Como método de referencia, se utilizó una tradicional bolsa de palabras (BoW por sus siglas en Inglés, Bag of Words). El esquema de pesado de términos utilizado corresponde a la frecuencia normalizada.

Vectores de palabras (word embeddings). Para explorar el uso de word embeddings, se empleó una representación basada en el promedio de vectores de palabras (AWE, por sus siglas en Inglés, Average Word Embeddings). Específicamente, cada tweet es representado por el promedio de los vectores de sus palabras. Para este propósito, se entrenaron word embeddings de 300 dimensiones usando la colección de letras misóginas a través del modelo Skip-gram.

En los experimentos se hace referencia a ellos como word embeddings especializados, ya que fueron generados específicamente para la tarea. Para fines de comparación, también se usaron embeddings generales previamente entrenados[5].

También se empleó una Unidad recurrente cerrada, GRU (por sus siglas en Inglés, (Gated Recurrent Unit) con una capa de atención. En este caso, los embeddings (especializados o generales) se utilizaron como la primera capa en el modelo.

Modelos del lenguaje pre-entrenados. Para los experimentos en Inglés y Español, se emplearon los modelos pre-entrenados distilbert-base-uncased [29] y BETO [11], respectivamente. Para la tarea multimodal, la representación textual de las frases se obtuvo aplicando Sentence-BERT [28] mediante el modelo all-MiniLM-L12-v1.

Todos los modelos fueron obtenidos de la librería de hugging-transformers[6]. También, se consideró un tamaño de lote (batch size) de 16 y la estrategia early stopping.

5.3. Clasificación y evaluación

Durante el proceso experimental, se utilizaron diferentes algoritmos de aprendizaje automático: Máquina de Vectores de Soporte (MVS), XGBoost (XG) [13] y Regresión Logística (RL). Como medidas de desempeño se reportaron la exactitud

[5] fasttext.cc/docs/en/pretrained-vectors.html
[6] huggingface.co/docs/transformers/index

Tabla 5. Evaluación de diferentes configuraciones en el proceso de filtrado: sin aumentación de datos (No), con aumentación de frases de canciones positivas y negativas, denotada como Frases, con frases provenientes de las técnicas de filtrado con instancias positivas (+), así como con instancias positivas y negativas (+)(-). Se reportan los valores de exactitud promedio, mínimo y máximo

Datos	Aumentación	Prom $\pm$ DE	Min	Max
Iber-Es	No	0.829 $\pm$ 0.015	0.810	0.854
	Frases	0.841 $\pm$ 0.011	0.826	0.859
	Coseno (+)	**0.851** $\pm$0.005	0.845	**0.859**
	Coseno (+) (-)	0.839 $\pm$ 0.010	0.828	0.857
	Roccio (+)	0.846 $\pm$ 0.004	0.846	0.856
	Roccio (+) (-)	0.844 $\pm$ 0.005	0.835	0.852
Iber-In	No	0.836 $\pm$ 0.010	0.822	0.853
	Frases	0.860 $\pm$ 0.016	0.844	0.886
	Coseno (+)	0.825 $\pm$ 0.013	0.810	0.842
	Coseno (+) (-)	0.861 $\pm$ 0.019	0.835	0.888
	Roccio (+)	0.843 $\pm$ 0.030	0.803	0.868
	Roccio (+) (-)	**0.892** $\pm$ 0.009	0.883	**0.906**
Eval-In	No	0.644 $\pm$ 0.018	0.617	0.671
	Frases	0.684 $\pm$ 0.010	0.666	0.694
	Coseno (+)	0.682 $\pm$ 0.016	0.658	0.705
	Coseno (+) (-)	**0.686** $\pm$ 0.016	0.663	**0.705**
	Roccio (+)	0.652 $\pm$ 0.004	0.645	0.656
	Roccio (+) (-)	0.666 $\pm$ 0.006	0.659	0.677

(Exac) y los valores F1. Los experimentos basados en el uso de redes neuronales y modelos pre-entrenados se realizaron cinco veces y se reportó el resultado promedio en la partición de prueba, así como la desviación estándar ($\mathrm{Prom} \pm \mathrm{DE}$).

6. Experimentos y resultados

En esta sección se reportan los experimentos y resultados que validan las ideas del método propuesto.

6.1. Evaluación de técnicas tradicionales

Como parte del trabajo experimental, se evaluaron dos técnicas comúnmente utilizadas para transferir conocimiento entre dominios: Adaptación de dominio y el uso de embeddings especializados (generados a partir de las letras de canciones etiquetadas como misóginas).

6.1.1. Adaptación de dominio

En términos generales, la adaptación de dominio (DA) busca entrenar un clasificador en un dominio y probarlo en otro que tiene una distribución de datos diferente [2].

En este sentido, el objetivo de este experimento es determinar la pertinencia de transferir el conocimiento lingüístico de las canciones como marcador de misoginia en el dominio de las redes sociales.

Para alcanzar este objetivo, se entrenaron clasificadores utilizando el contenido de las canciones, es decir, el dominio fuente, para clasificar instancias de conjuntos de datos de redes sociales, es decir, el dominio objetivo.

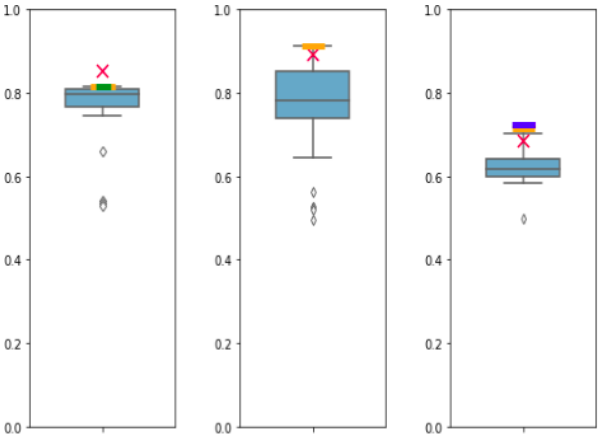


Fig. 5. Resultados en las respectivas tareas compartidas. Se ilustra la distribución de los valores de exactitud obtenidos en los conjuntos a) IberEval Español, b) IberEval Inglés, y c) Evalita Inglés. Las marcas de cruz en color rojo muestran el rendimiento alcanzado por la técnica propuesta

Particularmente, se entrenaron tres clasificadores mediante una BoW tradicional construida con las colecciones de datos: las letras completas de las canciones (denotada como LetraC) o con las frases de canciones etiquetadas (denotada como Frases). Como método de referencia se presentan los resultados de entrenar los clasificadores usando tweets en el entrenamiento y prueba (Tw).

La Tabla 3 presenta los resultados de este experimento. Se observa que el rendimiento de los clasificadores entrenados con las canciones completas o con las frases no superaron los resultados del método de referencia en las primeras dos colecciones.

Sin embargo, sí obtuvieron un mejor desempeño que un clasificador aleatorio en una tarea binaria, donde los resultados promedio se acercan al 50 %. Estos resultados indican la presencia de un subconjunto común de características en ambos dominios que tienen un valor relevante para detectar misoginia.

Además, de forma general, se observó que usando únicamente las frases de canciones se obtuvieron mejores resultados que empleando todas las canciones para el entrenamiento.

Este desempeño es notable en las colecciones Iber-Es e Iber-In. Estos resultados sugieren que los patrones lingüísticos que detectan misoginia están concentrados en algunas frases de las canciones y no en toda la letra. En general, los hallazgos encontrados pueden ser aprovechados para enriquecer enfoques y representaciones textuales más robustas.

6.1.2. Evaluación de embeddings

El objetivo de este experimento es evaluar métodos basados en word embeddings para transferir conocimiento de las letras de canciones hacia la tarea AMI. El interés es comparar el uso de embeddings especializados contra el uso de embeddings generales.

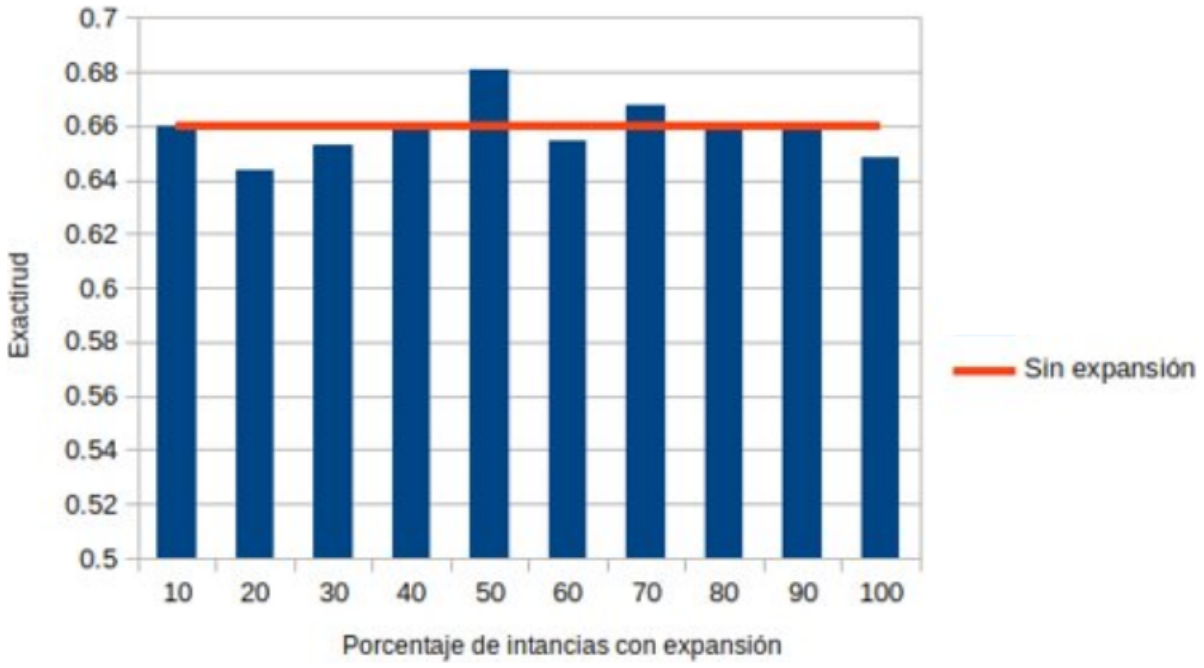
Los word embeddings especializados fueron aprendidos de las letras de las canciones que contienen contenido misógino explícito, por lo tanto, se generaron a partir de las canciones etiquetadas como misóginas. Mientras, los vectores generales corresponden a embeddings pre-entrenados de FastText.

Para llevar a cabo la comparación, se evaluaron dos representaciones basadas: i) el promedio de los embeddings individuales (AVE) clasificados por diferentes algoritmos de aprendizaje automático y ii) una arquitectura GRU alimentada por embeddings generales o particulares. La tabla 4 muestra los resultados de la comparación.

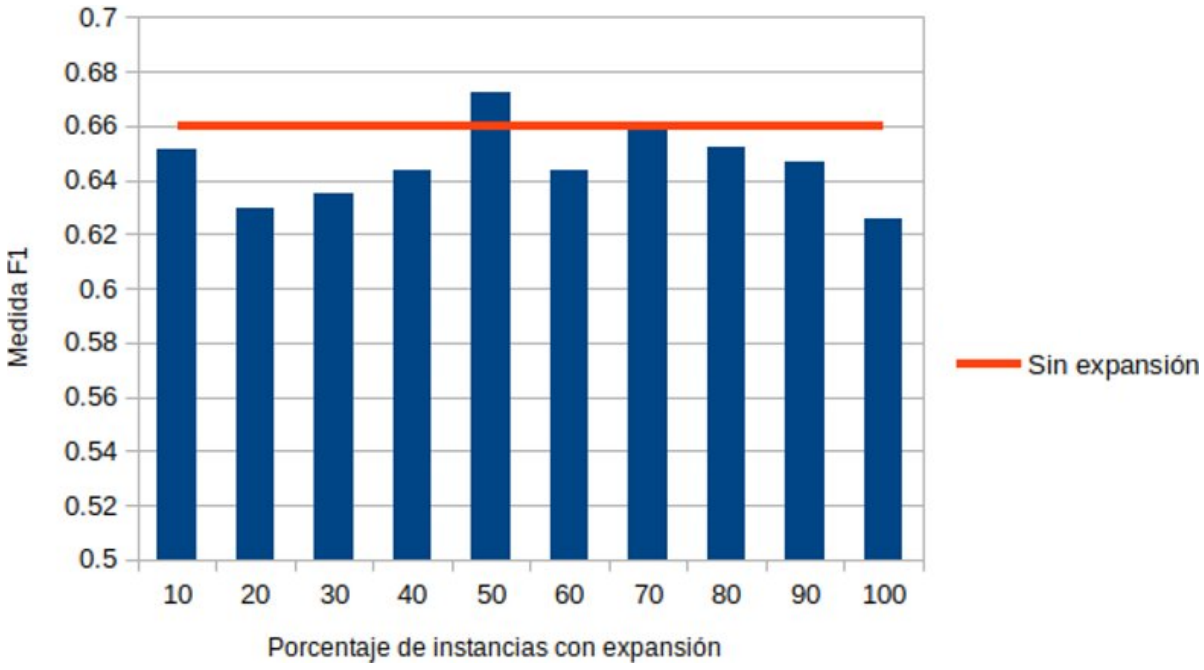
Los resultados mostraron que en la mayoría de los casos, los embeddings especializados se desempeñaron mejor que los embeddings generales, independientemente de la representación utilizada. No obstante, los resultados no superan el desempeño del BoW mostrada en el experimento anterior.

Este comportamiento indica que la representación textual es muy importante para aprovechar el conocimiento existente en las canciones y lograr transferirlo a la tarea AMI. Por lo tanto, es importante explorar nuevas representaciones que aprovechen el conocimiento contenido en las canciones.

ISSN 2007-9737



(a) Exactitud



(b) Valores F1

Fig. 6. Evaluación del método de expansión en la tarea de clasificación de memes. Se muestran los resultados obtenidos al aplicar la expansión del texto a diferentes porcentajes de instancias dentro de la colección de datos. Los resultados sin expansión son mostrados como método de referencia (línea roja)

6.2. Evaluación del enfoque propuesto para aumentación de datos

Los experimentos previos mostraron que el conocimiento de las canciones puede ayudar a detectar mensajes misóginos en las redes sociales. Esta sección está enfocada a evaluar el método de aumentación de datos propuesto, el cual tiene como objetivo diversificar los datos de entrenamiento en el dominio de destino (tweets) añadiendo ejemplos del dominio fuente (frases de canciones).

Dado que los modelos lingüísticos pre-entrenados (e.g., BERT) han mostrado resultados significativos en diferentes tareas de clasificación de textos, en este experimento se evaluaron modelos basados en BERT.

Particularmente, se utilizaron DistilBERT y Beto para los experimentos en Inglés y Español, respectivamente. Para fines de comparación, se evaluaron los mismos modelos con diferentes configuraciones en el entrenamiento, como se describe a continuación: sin aumentación de datos (No), con aumentación de frases originadas de canciones positivas y negativas (Frases), con frases provenientes de las técnicas de filtrado (filtrado basado en Similitud Coseno o basado el método Roccio) con instancias positivas (+) o con instancias positivas y negativas (+)(-). La Tabla 5 muestra los resultados del enfoque con las diversas configuraciones.

En general, los resultados muestran que todas las configuraciones de aumentación de datos obtuvieron mejor desempeño en comparación con los casos en los que no se aplicó aumentación. Por lo tanto, se concluye que las frases de las canciones son útiles para aumentar los datos de entrenamiento.

Además, es importante notar que el mejor resultado en cada conjunto de datos siempre fue obtenido con la técnica de aumentación de datos propuesta involucrando alguno de los mecanismos de filtrado (Coseno o Roccio).

Esto demuestra la utilidad de las frases de las canciones, pero sugiere una ventaja aún mayor cuando se utilizan únicamente aquellas frases de mayor calidad para la tarea.

Comparación con el estado del arte. En la Figura 5, el enfoque propuesto fue comparado con métodos del estado del arte para evaluar su competitividad. Primero, se contrastó con la distribución de resultados oficiales alcanzados en las tareas compartidas donde se han utilizado los conjuntos de datos empleados en este estudio.

Para fines de comparación, el desempeño del método de aumentación de datos propuesto se representa con cruces rojas y corresponde a las configuraciones que obtuvieron resultados los mejores resultados en cada conjunto de datos. Se pueden distinguir resultados competitivos en relación con los equipos participantes.

Es importante señalar que en el conjunto de datos Español, el método propuesto superó el desempeño del equipo ganador de la tarea compartida IberEval [18].

Sin embargo, en el conjunto de datos en Inglés, el rendimiento estuvo por debajo del ganador, lo cual ubicaría al método propuesto en el cuarto lugar de la competencia.

Por otro lado, en Evalita [17], los resultados obtenidos estuvieron ligeramente por debajo del primer lugar. El método propuesto también fue comparado con enfoques recientes y robustos del estado del arte que han usado los mismos conjuntos de datos.

En específico, en la figura, las marcas de color naranja corresponden a los resultados de un enfoque de clasificación de dominios cruzados que utiliza conjuntos de datos de diversas tareas relacionadas con el lenguaje abusivo, como discursos de odio y sexismo [25], empleando elementos del texto como emojis y clasificadores como MVS, GRU y BERT.

También se comparó con el rendimiento de un método de transferencia Bayesiano basado en una LSTM, el cual fue denotado con un guion azul [5]. Finalmente, se contrastó con resultados obtenidos por un método que usa una combinación de incrustaciones de palabras y características lingüísticas, el cual está representado con un guion verde [20]. En general, se demostró la competitividad del método frente a estos enfoques del estado del arte, resaltando su desempeño sobresaliente en el idioma Español.

6.3. Evaluación de la técnica de expansión de texto: detección de misoginia en memes

Este experimento está enfocado a abordar la tarea MAMI, una tarea multimodal enfocada a clasificar memes según la presencia de contenido misógino. En específico, el objetivo de este experimento es evaluar el impacto de la técnica de expansión del texto, la cual fue descrita en la Sección 4.

La idea de la técnica es expandir el texto de los memes con frases similares de canciones. En los experimentos, el texto fue expandido con la frase más similar dentro de la colección de frases etiquetadas generada en este trabajo de investigación. Una vez que el texto de las instancias es expandido, se entrena un clasificador multimodal.

Los resultados de este clasificador se muestran en las Figuras 6.a y 6.b, las cuales reportan los valores de exactitud y F1, respectivamente. Además, como método de referencia se muestra el desempeño obtenido por el modelo sin expansión del texto (línea roja en la figura).

Por otro lado, para analizar el comportamiento de acuerdo con el número de instancias en las cuales el texto ha sido expandido, se presentan los resultados obtenidos al aplicar la expansión en diferentes porcentajes de las instancias de entrenamiento. La figura muestra que el rendimiento del clasificador es mejorado cuando la expansión se realiza en el 50 % de las instancias del entrenamiento.

Estos resultados sugieren que el lenguaje de las canciones puede aumentar la diversidad lingüística de las expresiones textuales existentes en los memes. Sin embargo, es importante profundizar en el diseño de arquitecturas que tomen ventaja de estos hallazgos.

7. Conclusiones y trabajo futuro

En esta investigación se examinó la relevancia de las letras de las canciones para modelar manifestaciones de misoginia y transferir el conocimiento hacia la tarea de identificar de misoginia en redes sociales.

La idea que impulsó la investigación es la difusión de la ideología de género expuesta en una variedad de canciones, reflejando creencias socioculturales.

En particular, en este trabajo se propuso un método de aumentación de datos que aumenta la capacidad de generalización de los modelos de aprendizaje a través del uso de conocimiento proveniente de las canciones.

El enfoque fue evaluado en colecciones compuestas de publicaciones de redes sociales en el idioma Español e Inglés.

Los resultados experimentales mostraron que algunas canciones contienen patrones lingüísticos que reflejan manifestaciones misóginas y que este conocimiento puede ser transferido para detectar misoginia en contenido publicado en redes sociales.

Sin embargo, la riqueza de las canciones para este propósito se concentra únicamente en algunos fragmentos y no en toda la letra. Particularmente, los fragmentos relevantes pueden ser aprovechados para aumentar los datos de entrenamiento de la tarea a través de una evaluación de su calidad.

En este contexto, el enfoque propuesto superó los resultados del Estado del Arte en el idioma Español. El método puede ser adaptado para trabajar en escenarios de detección de misoginia multimodal mediante la expansión de los textos cortos con frases similares, brindando mayor información a los modelos computacionales.

Los resultados de esta investigación han motivado el interés de adaptar el método para trabajar con otros idiomas, por ejemplo, Italiano. Además, se planea diseñar arquitecturas y estrategias multimodales robustas que aprovechen el conocimiento de las letras de canciones.

Agradecimientos

Los autores agradecen al Consejo Nacional de Humanidades, Ciencias y Tecnologías de México (CONAHCYT) por financiar esta investigación mediante una beca para estudios de posgrado (CVU-714747).

Referencias

1. **Adams, T. M., Fuller, D. B. (2006).** The words have changed but the ideology remains the same: Misogynistic lyrics in rap music. Journal of Black Studies, Vol. 36, No. 6, pp. 938–957. DOI: 10.1177/0021934704274072.
2. **Alyafeai, Z., AlShaibani, M. S., Ahmad, I. (2019).** A survey on transfer learning in natural language processing. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp. 15–18. DOI: 10.18653/v1/N19-5004.
3. **Anzovino, M., Fersini, E., Rosso, P. (2018).** Automatic identification and classification of misogynistic language on twitter. Proceedings of the International Conference on Applications of Natural Language Processing and Information Systems, Vol. 10859, pp. 57–64. DOI: 10.1007/978-3-319-91947-8_6.
4. **Barton, G. (2018).** The relationship between music, culture, and society: Meaning in music. Music Learning and Teaching in Culturally and Socially Diverse Contexts: Implications for Classroom Practice, pp. 23–41. DOI: 10.1007/978-3-319-95408-0_2.
5. **Bashar, M. A., Nayak, R., Suzor, N. (2020).** Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. Knowledge and Information Systems, Vol. 62, No. 10, pp. 4029–4054. DOI: 10.1007/s10115-020-01481-0.
6. **Behm-Morawitz, E., Frisby, C. M. (2019).** Undressing the words: Prevalence of profanity, misogyny, violence, and gender role references in popular music from 2006-2016. Journal of Communication: Media Watch, Vol. 10, No. 1, pp. 5–21. DOI: 10.17613/7Y4W-JM88.
7. **Bicknell, J. (2002).** Can music convey semantic content? a Kantian approach. The Journal of Aesthetics and Art Criticism, Vol. 60, No. 3, pp. 253–261.
8. **Brook, B., Schindler-Zimmerman, T., Banning, J. H. (2008).** A feminist analysis of popular music. Journal of Feminist Family Therapy, Vol. 4, No. 18, pp. 29–51. DOI: 10.1300/J086v18n04_02.
9. **Calderón-Suarez, R. (2023).** Detección automática de contenido misógino en redes sociales mediante transferencia de conocimiento proveniente de canciones. Ph.D. thesis, Universidad Politécnica de Tulancingo.
10. **Calderón-Suarez, R., Ortega-Mendoza, R. M., Montes-Y-Gómez, M., Toxqui-Quitl, C., Márquez-Vera, M. A. (2023).** Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. IEEE Access, Vol. 11, pp. 13179–13190. DOI: 10.1109/ACCESS.2023.3242965.
11. **Canete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., Pérez, J. (2020).** Spanish pre-trained BERT model and evaluation data. Vol. 2020, pp. 1–10. DOI: 10.48550/arXiv.2308.02976.
12. **Chaudhury, S., Srivastava, K., Bhat, P., Sahu, S. (2017).** Misogyny, feminism, and sexual harassment. Industrial Psychiatry Journal, Vol. 26, No. 2, pp. 111. DOI: 10.4103/ipj.ipj_32_18.
13. **Chen, T., Guestrin, C. (2016).** XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining, pp. 785–794. DOI: 10.1145/2939672.2939785.
14. **Cooke, D. (1959).** The language of music. Oxford University Press.
15. **Farrell, T., Fernandez, M., Novotny, J., Alani, H. (2019).** Exploring misogyny across the manosphere in reddit. Proceedings of the 10th ACM Conference on Web Science, pp. 87–96. DOI: 10.1145/3292522.3326045.
16. **Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso,**

P., Lees, A., Sorensen, J. (2022). SemEval-2022 task 5: Multimedia automatic misogyny identification. Proceedings of the 16th International Workshop on Semantic Evaluation, pp. 533–549. DOI: 10.18653/v1/2022.semeval-1.74.

17. **Fersini, E., Nozza, D., Rosso, P. (2018).** Overview of the Evalita 2018 task on automatic misogyny identification (AMI). Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, pp. 59–66.
18. **Fersini, E., Rosso, P., Anzovino, M. (2018).** 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, Vol. 2150, pp. 214–228.
19. **Fischer, G. R. (1985).** How music communicates. Semiotica, Vol. 53, No. 1–3. DOI: 10.1515/semi.1985.53.1-3.131.
20. **García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R. (2021).** Detecting misogyny in spanish tweets. An approach based on linguistics features and word embeddings. Future Generation Computer Systems, Vol. 114, pp. 506–518. DOI: 10.1016/j.future.2020.08.032.
21. **Gourdine, R. M., Lemmons, B. P. (2011).** Perceptions of misogyny in hip hop and rap: What do the youths think?. Journal of Human Behavior in the Social Environment, Vol. 21, No. 1, pp. 57–72. DOI: 10.1080/10911359.2011.533576.
22. **Hewitt, S., Tiropanis, T., Bokhove, C. (2016).** The problem of identifying misogynist language on twitter (and other online social spaces). Proceedings of the 8th ACM Conference on Web Science, pp. 333–335. DOI: 10.1145/2908131.2908183.
23. **Li, B., Hou, Y., Che, W. (2022).** Data augmentation approaches in natural language processing: A survey. AI Open, Vol. 3, pp. 71–90. DOI: 10.1016/j.aiopen.2022.03.001.
24. **Lynn, T., Endo, P. T., Rosati, P., Silva, I., Santos, G. L., Ging, D. (2019).** A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary. Proceedings of the International Conference on Cyber Situational Awareness, Data Analytics And Assessment, pp. 1–8. DOI: 10.1109/CyberSA.2019.8899669.
25. **Pamungkas, E. W., Basile, V., Patti, V. (2020).** Misogyny detection in Twitter: A multilingual and cross-domain study. Information Processing and Management, Vol. 57, No. 6, pp. 102360. DOI: 10.1016/j.ipm.2020.102360.
26. **Peza-Casares, M. D. C. (2009).** Discurso de odio y feminicidios en méxico. Tram[p]as de la Comunicación y la Cultura, Vol. 66, pp. 29–35.
27. **Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., Martín-Valdivia, M. T. (2020).** Detecting misogyny and xenophobia in spanish tweets using language technologies. ACM Transactions on Internet Technology, Vol. 20, No. 2, pp. 1–19. DOI: 10.1145/3369869.
28. **Reimers, N., Gurevych, I. (2019).** Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.
29. **Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019).** DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. Proceedings of the 5th EMC2 - Energy Efficient Machine Learning and Cognitive Computing Colocated with the 33rd Conference on Neural Information Processing Systems, pp. 1–5. DOI: 10.48550/arXiv.1910.01108.
30. **Shushkevich, E., Cardiff, J. (2019).** Automatic misogyny detection in social media: A survey. Computación y Sistemas, Vol. 23, No. 4, pp. 1159–1164. DOI: 10.13053/CyS-23-4-3299.

31. Tomás, D., Ortega-Bueno, R., Zhang, G., Rosso, P., Schifanella, R. (2022). Transformer-based models for multimodal irony detection. Journal of Ambient Intelligence and Humanized Computing, Vol. 14, No. 6, pp. 7399–7410. DOI: 10.1007/s12652-022-04447-y.

32. Tsokalidou, R. (1989). Linguistic misogyny - a language universal: Observations, questions and ideas. Selected Papers on Theoretical and Applied Linguistics, Vol. 3, pp. 363–381. DOI: 10.26262/istal.v3i0.7182.

33. Weitzer, R., Kubrin, C. E. (2009). Misogyny in rap music: A content analysis of prevalence and meanings. Men and Masculinities, Vol. 12, No. 1, pp. 3–29. DOI: 10.1177/1097184x08327696.

34. Zeinert, P., Inie, N., Derczynski, L. (2021). Annotating online misogyny. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Vol. 1, pp. 3181–3197. DOI: 10.18653/v1/2021.acl-long.247.

Article received on 25/10/2023; accepted on 14/01/2024.
** Corresponding author is Félix Agustín Castro-Espinoza.*