LabTL-INAOE at MiSonGyny 2025: A Confidence-based Partitioning Strategy

Metztli Ramírez-González, Delia Irazú Hernández-Farías and Manuel Montes-y-Gómez

Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico

Abstract

In this article we describe the LabTL-INAOE participation in the MiSonGyny 2025 shared task. We present a method for detecting misogyny in song lyrics by a two-step classification framework based on a confidence-based partitioning strategy. In the first step, a base classifier is used to assign labels to all instances. These labels are then evaluated in order to assign them a confidence score. Based on this score, the instances are partitioned into two groups: trustworthy and untrustworthy. In the second step, the untrustworthy instances are relabeled using a more sophisticated model, in this case, a large language model with prompting. The proposed approach was evaluated in the two subtasks comprised in MiSonGyny 2025, yielding competitive results. Beyond quantitative performance, the method enhances explainability by providing confidence information at the instance level, making it especially useful for content moderation and other contexts where explainability and human oversight are essential.

Keywords

Classification ensemble, Prompting, Classification Confidence Score, Explainability

1. Introduction

Misogyny is defined as hateful behavior toward women, manifested in violent and cruel acts against them for the fact of being women [1]. Misogyny prevails in cultures and societies that consider women inferior to men and attribute to them a role centered on the reproduction of the human species, childcare, and homemaking. Because of this, women are exposed to physical violence, sexual abuse, degradation, unfair and humiliating treatment, as well as legal and economic discrimination.

This situation is fueled by the belief in the supposed inferiority of women and the overvaluation of male dominance, the latter being reinforced by factors such as traditionalism, family environment, and social media [1, 2]. Being so present in various aspects of everyday life, it is also reflected in music; the phenomenon of misogyny and sexism in song lyrics of any music genre is therefore a portrait of its population, both male and female, who assume and normalize violence against women [3].

There are studies that focus on analyzing specific genres to study the phenomenon of misogyny, for example, in genres such as hip-hop, rap, reggaeton, among other musical genres. [4, 5, 6]. This year, in the framework of IberLEF, the *MiSonGyny 2025* shared task was organized [7, 8] with the aim of detecting misogyny in Spanish Songs. The detection of misogyny in *MiSonGyny 2025* is divided into two tasks:

1. *Misogyny Speech Detection*: This task aims to classify phrases from song lyrics containing misogynistic speech. It is a binary taskin wich a song lyrics van belong to two labels: *a) Misogynist (M)* lyrics contain hate speech or disdain directed at women or perpetuate harmful gender stereotypes that promote subordination or objectification of women; *b) Not Misogynist (NM)* comprising lyrics that do not include hate speech or disdain against women. They may address themes related to women without perpetuating stereotypes or negative attitudes.
2. *Fine-grained Misogyny Speech Detection*: This task aims to predict the type of speech present in a phrase from a song. Tags are related with various types of hate speech related to misogyny:

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

metztli.ramirez@inaoep.mx (M. Ramírez-González); dirazuhf@inaoep.mx (D. I. Hernández-Farías); mmontesg@inaoep.mx (M. Montes-y-Gómez)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-10

- *Sexualization (S)*: Phrases that describe or suggest sexual acts, sexual language, or insinuations.
- *Violence (V)*: Phrases referring to physical or verbal aggression, threats, or violent actions.
- *Hate (H)*: Phrases containing offensive or discriminatory language, expressions of contempt, or hostility towards a group or individual.
- *Not Related (NR)*: Phrases that do not fall into the above categories and lack sexual, violent, or hateful content.

This paper describes our participation in the *MiSonGyny 2025* shared task; we developed a two-step classification method for detecting misogyny in song lyrics. In the first step, a base classifier is used to assign labels to all instances. These labels are then evaluated by a confidence-based partitioning strategy that computes a confidence score. Based on this score, the instances are divided into two groups: trustworthy and untrustworthy. In the second step, the untrustworthy instances are relabeled using a more sophisticated model, in this case, a large language model (LLM) with prompting.

This strategy is inspired by the saying *"different strokes for different folks"*. Some songs express misogyny explicitly, such as *"Run for Your Life"* by *The Beatles*: *"Well, I'd rather see you dead, little girl Than to be with another man"*. Other lyrics require deeper interpretation, such as *"Buenos días, amor"* by *José José*: *"Me perdí en tu vientre cuando aún dormías... Sé que estabas enfadada, pero no dijiste nada, El que calla otorga y sé que estás enamorada" (I was lost in your womb while you were still sleeping...I know you were angry, but you didn't say anything. Silence gives consent, and I know you're in love.)*. This excerpt implies a non-consensual sexual situation, though in a subtle manner. Finally, some lyrics require understanding of cultural and temporal context, such as *"There Goes My Everything"* by *Elvis Presley*: *"There goes my only possession, Oh, there goes my everything"* Although presented as a romantic sentiment, the idea of *"possession"* may be interpreted as inherently misogynistic.

As these examples show, each instance presents unique challenges in accurately detecting misogynistic content. Some can be handled by standard classifiers, while others demand more advanced models or even human judgment. Our proposed system adapts the analysis strategy according to the perceived difficulty of each instance.

This paper is organized as follows. In Section 2, we briefly introduce an overview of shared tasks related to identifying misogyny and hate speech in songs. In Section 3, we describe the experimental settings and the obtained results during the developing phase. In Section 4, we present the official results obtained in *MiSonGyny 2025* shared task. Finally, in Section 5, we conclude the paper.

2. Related Work

The identification of misogyny is part of the phenomena covered by *hate speech*, which is defined as a conscious and deliberate public statement intended to denigrate a group of people based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion or political affiliation [9].

Detecting hate speech is very challenging since it takes many forms in social media: it can be manifested verbally, non-verbally, and symbolically [10]. The music is presented as an organized language, a cultural component, and a generator of emotions. The types of violence presented in the lyrics of the different songs have been changing over time, although they seem to have a trend. The types of gender violence that involve domination through force have lost relevance, giving way to other more subtle forms of domination, such as symbolic and psychological violence, which have gained more strength [6]. For these reasons, the detection of hate speech in music, specifically the detection of misogyny is an open problem that must be approached with different solutions.

Within the field of NLP, several shared tasks have been proposed that provide labeled data and promote the development of proposals to solve different problems. To the present day, some tasks focused on the detection of misogyny or similar problems have been presented. Such as "Automatic Misogyny Identification" (AMI) in Evalita 2018 and Evalita 2020, focused on the identification of misogyny, the categorization of misogynistic behavior, and the classification of targets in tweets in Italian and English

[11, 12]. In the IberEval 2018 framework, an AMI task was also organized for the identification of misogyny, the categorization of misogynistic behavior, and the classification of the target of both Spanish and English tweets [13]. For IberLEF 2021 and 2022, "EXIST: sEXism Identification in Social Networks" was organized, a multilingual task for the identification and categorization of sexism in Spanish and English [14, 15]. And later, EXIST was organized for CLEF 2023 and CLEF 2024, also on the identification and categorization of sexism, but with others subtasks, one focused on source intention, and more subtasks with a multimodal approach for the identification and categorization of sexism in memes [16, 17].

Regarding tasks related to the detection of hate speech in songs, a subtask of HOMO-MEX 2024 was presented: Hate Speech Detection Towards the Mexican Spanish-speaking LGBT+ Population, where the HOMOLYRICS corpus was presented, composed of Spanish song lyrics that may or may not contain LGBT+phobic text. For this task, the main proposals were based on the use of traditional machine learning methods such as the Naive Bayes classifier and SVM classifier, others were focused on the use of prompting with LLMs such as Falcon or Llama 2, and the majority focused on the use of Transformers models and Transformers ensembles such as BERT, BETO, XLM RoBERTa, RoBerta, BERTweet, DistilBER, and mDeBERTa [18].

3. Experimental Methodology

In this work, we propose a two-step classification framework based on a confidence-based partitioning strategy. The main idea is to initially classify each test instance using a base model, selected for its simplicity or efficiency. Then, instead of accepting the predicted label as definitive, the model evaluates the confidence of that classification.

The **confidence score** is designed to reflect the attraction of a given instance to the most representative examples of the classes, both, the predicted class and the opposing class. If an instance is strongly attracted to the core examples of its predicted class and only weakly attracted to those of the opposite class, then the confidence in the prediction should be considered high.

To compute this score, we adopt a strategy inspired by the K-Strongest Strengths (KSS) algorithm [19]. This approach models the idea of attraction forces between texts, where each instance exerts a force on another based on their semantic similarity and relevance (masses). The mass of an example is derived from its importance within the training data; in this case, we use the cumulative frequency of class-relevant words. By combining these elements, we compute the average attraction force exerted on the instance by both the predicted class and the opposing class.

The resulting **confidence score** is obtained by calculating the ratio between the average attraction force from the predicted class and that from the opposing class. This ratio determines whether an instance is *trustworthy* (i.e., classified with confidence) or *untrustworthy* (i.e., uncertain or subtle). Difficult instances are passed to a more sophisticated and costly model (e.g., a GPT-based classifier via prompting) for relabeling.

3.1. Our Method

Our method consists of four steps, which are illustrated in Figure 1. This framework is particularly well-suited for subjective classification tasks, where the boundaries between classes are not always clearly defined, such as the detection of misogyny in song lyrics.

1. **Songs Data:** The data includes song phrases of varying lengths, from a variety of genres, and covering a wide range of topics. They are divided into training and test sets for both tasks.
2. **Preprocessing:** Instances are standardized by converting all text to lowercase, removing line breaks, punctuation marks, and structural indicators such as choir, verse, etc.
3. **First classifier:** The first model used to assign labels is RoBERTuito, which was fine-tuned for this task. From this model, we obtain both the predicted labels to be later evaluated, and the

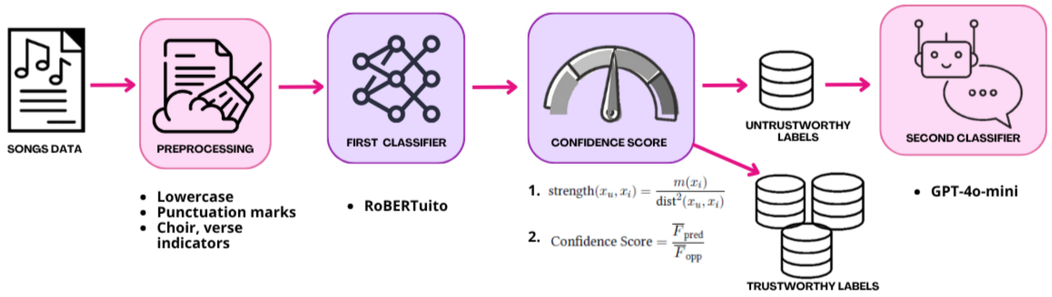
SONGS DATA
PREPROCESSING
FIRST CLASSIFIER
CONFIDENCE SCORE
UNTRUSTWORTHY LABELS
SECOND CLASSIFIER
TRUSTWORTHY LABELS
• Lowercase
• Punctuation marks
• Choir, verse indicators
• RoBERTuito
1. $\text{strength}(x_u, x_i) = \frac{m(x_i)}{\text{dist}^2(x_u, x_i)}$
2. $\text{Confidence Score} = \frac{\overline{F}_{\text{pred}}}{\overline{F}_{\text{opp}}}$
• GPT-4o-mini

Figure 1: Diagram of the proposed approach.

embeddings, which serve as the base representation of the song lyrics and are used to compute the confidence score. [20].

4. **Confidence score:** This process is inspired by the kSS algorithm (k-strongest strengths classification algorithm), which draws an analogy with Newton's Law of Universal Gravitation. Unlike traditional approaches that rely solely on the labels of neighboring instances, kSS uses the gravitational forces exerted by training instances to assess their influence [19].
   Following this principle, we construct the attraction forces for each test instance. Using the word embeddings from the Transformer model, we calculate the cosine distances between the training and test objects.
   To assign a mass to each training instance, we rely on the relevance of specific n-grams for each class. These n-grams are extracted using the SS3 algorithm [21], which identifies the features that contribute the most to distinguishing between classes. For each class, we construct a list of the most informative n-grams. Then, for every training instance, we count the frequency of the n-grams from the list corresponding to its predicted class. The resulting count determines the instance's mass: the higher the frequency of class-relevant n-grams it contains, the greater its mass. Finally, the attraction force it exerts on a test instance is computed using the relationship between the previously calculated mass and the cosine distance, as defined in Formula 1.

   $$\text{strength}(x_u, x_i) = \frac{m(x_i)}{\text{dist}^2(x_u, x_i)} \tag{1}$$

   For each class, the 10 strongest forces acting on the test instance are selected to calculate the confidence score. we consider the predicted label assigned by the base model and the distribution of forces. As shown in Formula 2, we calculate the average force from the predicted class and divide it by the average force from the opposing class(es). The result is a single value representing the **confidence score** of the instance. Instances with scores above a given threshold (typically between 1.0 and 1.5) are considered *trustworthy* and retain their assigned label. Those with lower scores are marked as *untrustworthy*, as they exhibit traits associated with multiple classes, making them challenging to classify automatically.

   $$\textit{Confidence Score} = \frac{\overline{F}_{\text{pred}}}{\overline{F}_{\text{opp}}} \tag{2}$$

5. **Second classifier:** Instances deemed *untrustworthy* (i.e., with low confidence scores) are passed to a large language model (LLM), specifically GPT-4o-mini [22], for re-labeling via prompting (the following section shows the prompts used). The final labels for this subset are those produced by the LLM.

4. Official Results and Results Analysis

4.1. Dataset

For training purposes, task organizers provided a dataset for each subtask:

- Task 1: It has a total of 2104 training data, divided into 642 instances for the misogyny class, and 1462 instances for the non-misogyny class. The test data has a total of 527 instances.
- Task 2: It has a total of 1168 training instances, divided into 435 instances marked as Sexualization(s), 129 instances marked as Violence (V), 78 instances marked as Hate (H), and 526 instances marked as Not Related (NR). The test data has a total of 293 instances.

4.2. Results

Table 1 presents the results of the internal validation conducted on the training data for both tasks. This evaluation compares the performance of a basic method (Bag of Words) and the original Transformer model with our proposed confidence-based partitioning strategy. Table 1 reports the number of instances in each partition, the F1-score obtained for trustworthy and untrustworthy instances. For both tasks, the dataset was split into 80% for training and 20% for testing. Specifically, Task 1 consisted of 1,683 training instances and 421 testing instances, while Task 2 included 934 training instances and 234 testing instances.

The *trustworthy* labels achieved a high performance, while *untrustworthy* labels yielded a significantly lower performance in both tasks. This clear contrast in performance between the two subsets suggests that the partitioning strategy effectively distinguishes between explicit and implicit examples. The drop in performance on hard instances is expected and desirable, as it validates the model's ability to identify those cases where classification is inherently more challenging.

Table 1
Evaluation for the First and Second Task in the training set.

Task	Classifier	Instances	%	F1
Task 1	BoW	421	100	0.7263
	RoBERTuito	421	100	0.7988
	Trustworthy	287	68.17	0.9284
	Untrustworthy	134	31.83	0.4826
Task 2	BoW	234	100	0.3696
	RoBERTuito	234	100	0.4713
	Trustworthy	195	83.33	0.6185
	Untrustworthy	39	16.67	0.32

Based on the results obtained during system validation, the proposed method was applied to the test partition, separating the instances according to their confidence level. The results of this separation for each task are described below:

- Task 1:The method identified 102 instances (19.35%) as *untrustworthy*, out of a total of 527. These were labeled using prompting with the model GPT-4o-mini, using a specific prompt designed for this task.

> "Eres un modelo de clasificación de canciones misóginas que asigna una categoría a cada oración, esto tomando en cuenta que la misoginia se define como el odio o prejuicio contra las mujeres, puede manifestarse lingüísticamente de numerosas maneras, incluida la exclusión social, la discriminación, la hostilidad, las amenazas de violencia y la cosificación sexual. Las clases posibles para clasificar son: 'Misogino', 'No Misogino'. Responde solo con la etiqueta exacta, 0 para No Misogino, 1 para Misogino."
>
> ---
>
> *"You are a misogynistic song classification model that assigns a category to each sentence. This is based on the definition of misogyny as hatred or prejudice against women, which can be linguistically expressed in numerous ways, including social exclusion, discrimination, hostility, threats of violence, and sexual objectification. The possible classes for classification are: 'Misogynistic', 'Not Misogynistic'. Respond only with the exact label: 0 for Not Misogynistic, 1 for Misogynistic."*

- Task 2: 58 instances (19.80%) were detected as *untrustworthy*, out of a total of 239, also labeled with GPT-4o-mini using the same approach. The prompt used is shown below:

> "Eres un modelo de clasificación de canciones misóginas que asigna una categoría a cada oración, esto tomando en cuenta que la misoginia se define como el odio o prejuicio contra las mujeres, puede manifestarse lingüísticamente de numerosas maneras, incluida la exclusión social, la discriminación, la hostilidad, las amenazas de violencia y la cosificación sexual. Sexualización (S) : Frases que describen o sugieren actos sexuales, lenguaje sexual o insinuaciones. Violencia (V) : Frases que se refieren a agresiones físicas o verbales, amenazas o acciones violentas. Odio (H) : Frases que contienen lenguaje ofensivo o discriminatorio, expresiones de desprecio u hostilidad hacia un grupo o individuo. No relacionado (NR) : Frases que no entran en las categorías anteriores y carecen de contenido sexual, violento o de odio.. Responde solo con la etiqueta exacta, S, V, H o NR."
>
> ---
>
> *"You are a misogynistic song classification model that assigns a category to each sentence, taking into account that misogyny is defined as hatred or prejudice against women, it can manifest itself linguistically in numerous ways, including social exclusion, discrimination, hostility, threats of violence and sexual objectification. Sexualization (S): Phrases that describe or suggest sexual acts, sexual language or innuendos. Violence (V): Phrases that refer to physical or verbal aggression, threats or violent actions. Hate (H): Phrases that contain offensive or discriminatory language, expressions of contempt or hostility towards a group or individual. Unrelated (NR): Phrases that do not fall into the previous categories and lack sexual, violent or hateful content. Answer only with the exact label, S, V, H or NR."*

For both tasks, the final labels were generated by joining the preditions on the two subsets, that is:

- The labels obtained with the fine-tuned RoBERTuito model for the easy instances.
- The labels generated by prompting with GPT-4o-mini for the hard instances.

The results obtained are summarized in Table 2. As can be seen, the performance values remain practically unchanged compared to the use of the base model, indicating that the separation strategy does not significantly improve the classification. While our method achieves a slight improvement in Task 1, there is no difference in Task 2. This is mainly due to the fact that Task 2 poses additional challenges, as it is a multi-class problem with some classes having very few instances, which limits their representativeness and reduces the effectiveness of the proposed approach.

Table 2
Official Results for the First and Second Task.

Task	Classifier	F1	Ranking
Task 1	RoBERTuito	0.7724	-
	Our Method	0.7833	eighth
Task 2	RoBERTuito	0.4226	ninth
	Our Method	0.4226	-

4.3. Results Analysis

Although the proposed approach does not lead to a substantial increase in quantitative performance metrics, it provides significant qualitative advantages. In particular, it enhances the explainability of the classification process and offers valuable support for human decision-making. This is especially relevant in sensitive applications such as content moderation, where the ability to understand and justify predictions is as important as their accuracy. Rather than aiming solely for metric improvement, the method focuses on enriching the interpretability of results, enabling more informed and transparent decisions.

4.3.1. Highlights of the method

- **Lists of relevant n-grams:** For each class, the system generates a set of representative n-grams that are used to calculate the masses within the attraction model. Table 3 shows the most significant n-grams that characterize each category.

Table 3
Examples of n-grams for the classes in the first task.

Classes	Uni-grams	Bi-grams
Misogynist	baby, mami, flow, vamo, encima, cama, culo, novio, mueve, sexo, party, mujeres, perra, perreo, tocarte.	si quiere, daddt yankee, j balvin, pal carajo.
	baby, mommy, flow, let's go, on top, bed, ass, boyfriend, move, sex, party, women, bitch, twerking, touching yourself.	*If you want, daddy yankee, j balvin, fuck off.*
Not Misogynist	alma, vivir, dolor, jamás, luz, fin, adiós, triste, amar, soledad, llorar, silencio, sufrir, cariño, morir.	tantas cosas, pido perdón, ser feliz, si pudiera, solo queda, tal vez.
	soul, live, pain, never, light, end, goodbye, sad, love, loneliness, cry, silence, suffer, affection, die.	*So many things, I apologize, to be happy, if I could, all that remains is, maybe.*

- **Trust ranking per instance:** Each song is given a confidence score that allows it to be ranked from most to least reliable. This ranking is useful for prioritizing subtle cases or candidates for *human review*. The system automatically identifies a small subset of sentences (around 20 %) that contain subtle forms of sexist language, which can be reviewed with more sophisticated tools or by human moderators.

To demonstrate the usefulness of the method, below we show concrete examples that were marked as *trustworthy* or *untrustworthy*. Table 4 presents examples of sentences classified as **trustworthy**. The misogynistic examples include clear expressions of sexualization and objectification. In contrast, sentences classified as not misogynistic, although they may contain affective or animated language, do not have evidence of hateful or prejudiced content.

On the other hand, Table 5 shows examples of sentences labeled as misogynistic by GPT-4o-mini, initially considered implicit or **untrustworthy** by the model. In the first case, metaphors of physical violence are detected. In the second, there is a subtle allusion to female sexual freedom, which can be interpreted in different ways, some may attribute and even explain the use of the adjective to the person's infidelity, whereas others may notice a clear misogynistic expression, which shows that many

Table 4
High-confidence instances.

	Example 1	Example 2
Misogynist	Eres una hija 'e puta en la cama que hace posiciones solo para mí Siempre que te me pongas bellaca, sabes que estoy pa' ti.	Yo te vo'a poner en cuatro y te vo'a pose-e-er (¡Jaja!) Y como una bala en un muerto, yo te penetré, eh-eh
	You're a bitch in bed *who does positions just for me* *Whenever you get nasty with me,* *you know I'm here for you*	*I'm going to put you on all fours and* *I'm going to poss-e-s-s you (Haha!)* *And like a bullet in a dead man,* *I penetrated you, eh-eh*
Not Misogynist	Calienta esto, que esto está frío Calienta esto, que esto está frío Cuando yo llego a tocar me piden los amigos míos.	La costumbre se hace triza Y al final lo he comprobado Tanto como un loco Te quiero a mi lado
	Heat this up, it's cold. *Heat this up, it's cold.* *When I get to play,* *my friends ask me.*	*Habit is shattered* *And in the end I've proven it* *As much as a madman* *I want you by my side*

of these examples end up being so subtle that automatic methods could not perfectly identify them. These situations reflect the complexity of language and the need for more refined analysis mechanisms or human intervention.

Table 5
Subtle instances of low confidence.

	Example 1	Example 2
Subtle	Con la punta'e palo (le voy a dar) Con el medio palo (pa' que aprenda a respetar) Con el palo entero (que conmigo no se juega) Con la punta'e palo (y la vuelvo a castigar) Con el medio palo (yo la vuelvo a rematar)	Mira estos celos, me están matando Ay, mujer, que fácil eres Abres tus alitas, muslos de colores Donde se posan tus amores Mariposa traicionera
	With the tip of the stick (I'm going to hit her) *With the half stick (so she learns to respect)* *With the whole stick (she can't play with me)* *With the tip of the stick (and I punish her again)* *With the half post (I finish her again)*	*Look at this jealousy, it's killing me.* *Oh, woman, how easy you are.* *You open your little wings, colorful thighs* *Where your loves rest* *Treacherous Butterfly*

5. Conclusions

The proposed system introduces a strategy for performing a confidence assessment that allows instances to be divided into *trustworthy* and *untrustworthy* subsets. This division allows efficient models like RoBERTuito to be applied to easy instances and more powerful models like GPT-4o-mini to be reserved for the hard instances that require a more in-depth analysis. Although the quantitative results do not show a significant improvement in classification metrics, the added value of the system lies in its ability to explain the decision-making process, identify challenges and provide useful tools for content moderation.

The system's main contributions include: (i) the identification of relevant n-grams for each class, (ii) the calculation of an interpretable confidence score for each instance, and (iii) the ability to generate a reduced subset of implicit examples as candidates for human review. These features make the system not only function as a classifier, but also as a tool for analyzing and interpreting textual content, particularly valuable in sensitive contexts such as detecting misogyny in song lyrics.

In summary, this work contributes a hybrid approach that prioritizes explainability and efficiency in the use of computational resources, opening new possibilities for the design of moderation systems that are fairer, more understandable, and adaptable to complex social contexts.

Declaration on Generative AI

Generative AI tools were used solely within the proposed classification approach to assist in the processing of certain instances. Grammarly was employed to suggest minor grammatical and stylistic corrections. No generative AI systems were used to write or edit the content of this manuscript.

References

[1] Instituto Nacional de las Mujeres, Misoginia, n.d. URL: https://campusgenero.inmujeres.gob.mx/glosario/terminos/misoginia, recuperado el 29 de mayo de 2025.

[2] K. Galarza, Misoginia, odio hacia las mujeres, n.d. URL: https://antares.iztacala.unam.mx/pieg/index.php/articulos-gaceta/violencia/misoginia-odio-hacia-las-mujeres/, recuperado el 29 de mayo de 2025.

[3] N. Contreras, La misoginia en la música, https://radio.uabc.mx/podcast/la-misoginia-en-la-musica, 2022. Podcast producido por UABC Radio.

[4] Y. Xian, Gender inequality and misogyny in hip-hop music, in: SHS Web of Conferences, volume 193, EDP Sciences, 2024, p. 02018.

[5] J. Q. K. Ling, G. F. Dipolog-Ubanan, Misogyny in the lyrics of billboard's top rap airplay artists, International Journal of Arts Humanities and Social Science 2 (2017) 7–13.

[6] A. Alpízar-Lorenzo, L. Hernández-Muñoz, M. C. Ledezma-Trujillo, M. A. Linares-Villa, L. Rodríguez-Cortés, G. Guzmán-Díaz, J. Cisneros Herrera, Goce sin límites: manifestaciones misóginas en canciones de reggaetón, Boletín Científico De La Escuela Superior Atotonilco De Tula 8 (2021) 25–32. URL: https://doi.org/10.29057/esat.v8i15.6510. doi:10.29057/esat.v8i15.6510.

[7] T. Alcántara, M. Soto, C. Macias, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveron, Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural 75 (2025).

[8] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[9] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, H. M. H. López, Internet, social media and online hate speech: Systematic review, Aggression and Violent Behavior 58 (2021) 101608. doi:10.1016/j.avb.2021.101608, art. no. 101608.

[10] M. A. Paz, J. Montero-Díaz, A. Moreno-Delgado, Hate speech: A systematized review, Sage Open 10 (2020) 2158244020973022. doi:10.1177/2158244020973022, art. no. 2158244020973022.

[11] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), volume 2263, CEUR-WS.org, 2018, pp. 1–9. URL: https://ceur-ws.org/Vol-2263/paper009.pdf.

[12] E. Fersini, D. Nozza, P. Rosso, Ami@evalita2020: Automatic misogyny identification, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), volume 2765, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/paper161.pdf.

[13] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: Proceedings of IberEval 2018, volume 2150, CEUR-WS.org, 2018, pp. 214–228. URL: https://ceur-ws.org/Vol-2150/overview-AMI.pdf.

[14] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: Sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: https://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[15] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, P. Rosso, Overview of exist 2022: Sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: https://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[16] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: Sexism identification in social networks, in: Proceedings of the 45th European Conference on Information Retrieval (ECIR 2023), Springer, 2023, pp. 593–599. URL: https://doi.org/10.1007/978-3-031-28241-6_68. doi:10.1007/978-3-031-28241-6_68.

[17] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, D. Spina, Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes, in: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2024), Springer, 2024, pp. 93–117. URL: https://ceur-ws.org/Vol-3740/paper-87.pdf.

[18] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macías, Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.

[19] J. Aguilera, L. C. González, M. Montes y Gómez, R. López, H. J. Escalante, From neighbors to strengths - the k-strongest strengths (kss) classification algorithm, Pattern Recognition Letters 136 (2020) 301–308.

[20] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[21] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, T-ss3: A text classifier with dynamic n-grams for early risk detection over text streams, Pattern Recognition Letters 138 (2020) 130–137. doi:10.1016/j.patrec.2020.07.019.

[22] OpenAI, Gpt-4o-mini [large language model], 2024. URL: https://chat.openai.com/, accessed: 2025-05-14.