CEUR-WS.org/Vol-4098/MiSonGyny2025_paper4.pdf

PRHLT-UPV at MiSonGyny 2025: Detection of Misogynist Expressions in Song Lyrics

Elias Urios Alacreu[1,*,†], Maria Aloy Mayo[1,†] and Paolo Rosso[1,2]

[1]*PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain*
[2]*ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain*

Abstract

Gender bias and discrimination against women are still manifested in popular culture through various types of discourse present in social media, TV shows, and music. In particular, song lyrics often express misogyny in explicit ways (e.g., objectification of women's bodies, violent threats), but also in covert forms masked as romantic expression. This paper reports our participation in the MiSonGyny 2025 shared task, which involves both the detection and fine-grained classification of misogynistic content in lyrics. Our experiments explore the effectiveness of BERT-like models and large language models (LLMs) for these tasks.

Keywords

Misogyny detection, song lyrics, gender bias, natural language processing, LLMs

1. Introduction

The spread of misogynistic discourse remains a persistent issue contributing to gender inequality. Harmful content towards women in social media platforms and other ways of communication implies a critical scenario that promotes gender inequality. It also contributes "to an environment where women feel unsafe, targeted or silenced"[1].

We can consider misogyny within the concept of hate speech (HS) understood as "any communication denigrating a person or group, judging by specific features such as color, ethnicity, gender, sexual orientation, nationality, and religion" [2]. In this sense, Moreno-Sandoval et al. consider misogynistic speech "when hate speech is gender-oriented and specifically targets women" [3]. Moreover, social studies link misogynistic discourses with the enforcement of male dominance, since these attitudes promote discrimination against women and "its effects are widespread and systematic, bearing severe both social and individual consequences, such as verbal and physical violence, rape and femicide" [4]. In summary, the presence of misogyny in social media and popular culture discourses denotes hateful attitudes or actions against women that still prevail in society.

This scenario presents a challenge for research studies trying to understand the semantic patterns hidden beneath a wide range of written and visual features in discourses such as social media platforms, news articles or TV shows. For instance, even song lyrics are contaminated with ideologically-biased ideas based on misogynistic attitudes, and we believe that it is crucial to address this concern, given the influential role music plays in shaping societal perspectives. The powerful influence of music is raised in Davies [5] emphasizing that song lyrics "are more than mere mirrors of society; they are a potent force in the shaping of it". This idea reveals the power of linguistic structures present in song lyrics and how the application of Natural Language Processing (NLP) techniques presents unique challenges. Factors such as the use of metaphors, slangs, and poetic language can obscure explicit content and make misogyny in music harder to detect for Natural Language Models. Addressing these complexities requires the development of specialized models capable of interpreting nuanced language and cultural references inherent in musical texts.

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

elural2@prhlt.upv.es (E. U. Alacreu); malomay@upvnet.upv.es (M. A. Mayo); prosso@dsic.upv.es (P. Rosso)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

published 2025-12-10

In this regard, an interesting work by Betti et al. [6] analyses the 'Two Million Song Database' corpus, focusing on the detection of sexism over five decades (1960-2010). The study employed NLP techniques to identify sexist language and measured gender biases through word embeddings, highlighting the prevalence of misogynistic themes in popular music. Their conclusions reveal an increase in sexist content and specifically observe that "songs of male solo artists contain more and often stronger biases than those of female solo artists". It shows how lyrics carry strong cultural and contextual signals and are the expression of popular culture that can influence how listeners see the world and reflect how artists perceive it [6]. Their conclusions highlight the continuing need to explore gender bias in popular music, focusing not only on linguistic structures but also on other sociocultural aspects such as the gender of the artist or even the impact on the audience.

This scenario confirms that the exploration of misogyny in song lyrics has opened new avenues for multidisciplinary research, and it is a call to action highlighting the need for models attuned to the intricacies of lyrical language. Continued efforts in dataset development, bias mitigation, and model refinement are essential to further progress in this critical area of study.

The following work aims to investigate this gap by participating in the MySonGyny shared task [7] at IberLEF 2025 [8], which proposes both the detection of misogyny in song lyrics as well as a more fine-grained classification. Our system presents a comparison of BERT-like models against LLM models to detect mysogyny speech in song lyrics. In addition, we also investigate the effects of data augmentation by LLM for the given task.

2. Related Work

In recent years, the proliferation of HS on digital platforms has sparked significant interest within the NLP community. Numerous resources, shared tasks, and benchmark datasets [9, 10, 11, 12, 13, 14, 15, 16] have been proposed, contributing to the development of classifiers capable of effectively addressing this challenge.

Misogyny, as a specific form of HS, has received particular attention. Existing studies have tackled this issue primarily through the textual analysis of comments extracted from social media platforms, especially X [9, 10, 11, 12, 13, 14]. More recently, research has expanded to multimodal approaches, including the detection of misogyny in memes [15, 12].

Despite these advances, most approaches still rely heavily on social media data, which limits the generalization of models to other forms of communication. Creative and artistic texts, such as song lyrics, present unique challenges for the automatic detection of HS. Although some recent work has addressed this issue [6, 17, 18], the number of shared tasks dedicated to this topic remains small, with only the third subtask of HOMO-MEX 2024 focusing on LGBTQ+phobia detection in song lyrics [19].

From a methodological standpoint, recent advances in Deep Learning (DL) have significantly shaped the development of HS detection systems, with Transformer-based architectures [20] setting the state of the art over the last few years. On the one hand, Transformer encoder models [21, 22, 23] have become standard due to their ability to produce contextual representations and adapt through fine-tuning. On the other hand, Large Language Models (LLMs), such as GPT [24] or LLaMA [25], are generative models based on Transformer decoder architectures and trained on massive corpora. These models have demonstrated strong performance in classification tasks [26] by leveraging few-shot learning [27] capabilities and advanced reasoning skills through techniques such as Chain of Thought (CoT) prompting [28].

3. Tasks and Dataset Descriptions

3.1. Tasks Descriptions

The MiSonGyny 2025 shared task [7] is divided into two sub-tasks:

1. **Misogyny Speech Detection**. A binary classification task aimed at classifying phrases from song lyrics containing misogynistic speech. The classes are described as follows:
    - **Misogynist (M)**: Any lyrics that contain hate speech or disdain directed at women or perpetuate harmful gender stereotypes that promote subordination or objectification of women.
    - **Not Misogynist (NM)**: Lyrics that do not include hate speech or disdain against women. Although they might address themes related to women, they can do so without perpetuating stereotypes or negative attitudes.
2. **Fine-grained Misogyny Speech Detection**. A hierarchical multi-class classification task aimed at detecting the type of speech present in a song. The categories can be described as follows:
    - **Sexualization (S)**: Phrases which describe or suggest sexual acts, sexual language, or insinuations.
    - **Violence (V)**: Refers to lyrics containing physical or verbal aggressions, threats or violent actions.
    - **Hate (H)**: Lyrics which contain offensive or discriminatory language, expressions of contempt, or hostility towards a women or a group of women.
    - **Not Related (NR)**: Phrases that do not fall into the above categories and lack sexual, violent, or hateful content.

Figure 1 depicts a visual overview of these tasks and their relationships.

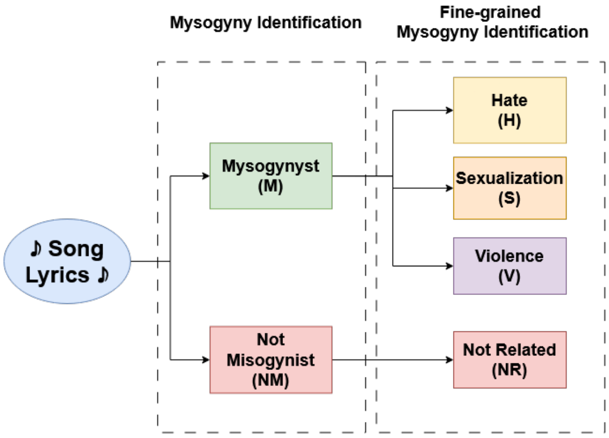


Figure 1: Overview of the MiSonGyny 2025 shared task structure, showing the relationship between the Misogyny Speech Detection sub-task and the hierarchical Fine-grained Misogyny Speech Detection task categories.

For the evaluation of the proposed systems across both sub-tasks, standard classification metrics such as Precision (Pr), Recall (Rc) and F1-score were proposed by the organizers, with F1-score being the selected metric for determining the ranking on the official test set [7]. In addition to the proposed metrics, our system's performance is also evaluated with the Area Under the Receiver Operating Characteristic Curve (AUROC) [29] [1] as well as the Matthews Correlation Coefficient (MCC) [30], which provides us with a deeper understanding of the behaviour of our system.

Table 1 summarizes the metrics used for the evaluation of each sub-task, indicating the official metrics used for ranking.

3.2. Dataset

This section describes the dataset used in the MiSonGyny 2025 shared task [7]. The dataset distributions for each sub-task are illustrated in Figure 2.

---

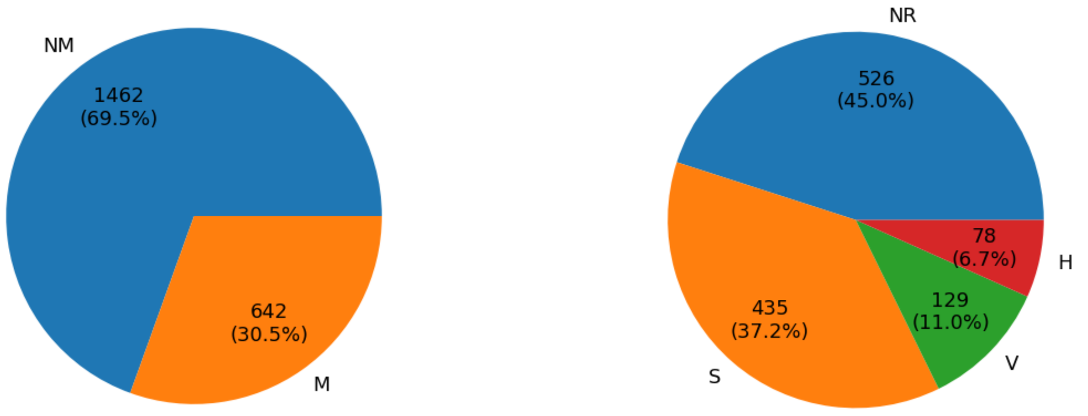
[1] AUROC will not be present in the evaluation of all our system since it requires class probabilities.

Subtask	Metrics
1	AUROC, MCC, **F1**, Pr, Rc
2	AUROC, MCC, **F1**, Pr, Rc

Table 1
Metrics used for the evaluation of both sub-tasks. Bolded metrics are the ones used for the official evaluation during the competition.

On the one hand, Figure 2a presents the data distribution for the first subtask. Although we are provided with approximately 2300 samples, it is important to note on the class imbalance present for this sub-task, since the non-misogynistic songs make up almost 70% of the dataset.

On the other hand, Figure 2b presents the data distribution for the second subtask. Since this subtask is a hierarchical one, all the samples on this subtask are directly extracted from the previous. Likewise, the most common class is the one related to the absence of misogyny (i.e., NR for this subtask). By examining the categories related to the presence of hate speech, it can be observed that the sexualization class is the most common among them with 435 samples, whereas the violence and hate classes only make up less than 20% of the overall dataset, presenting a very skewed distribution.



(a) Misogyny Speech Detection sub-task.

(b) Fine-grained Misogyny Speech Detection sub-task.

Figure 2: Dataset distribution for each of the sub-tasks of MySonGyny.

4. System Proposal

4.1. Preprocessing

Song lyrics possess a unique structure, driven by their domain, that requires specific NLP considerations. For instance, they frequently feature pervasive repetition, like recurring choruses and verses, which, while integral to musical composition, offers little to no information for the given tasks.

Beyond repetition, lyrics often contain elements irrelevant to their core message. This includes ad-libs (often sounds rather than coherent sentences) and song delimiters such as [Intro] or [Chorus]. As shown in Table 2, these elements, coupled with the initial song length, frequently push them beyond common token limits like BERT's 512-token maximum.

Believing these structural and superfluous elements do not significantly influence hate speech presence, we preprocessed our data by removing ad-libs, song section delimiters, blank lines, and repeated verses and sections. As Table 2 reflects, this step allows approximately 75% of the songs to fit within the BERT model's token limit, enabling more efficient and relevant analysis focused on the unique and semantically relevant lyrical content.

Table 2
Token length statistics of song lyrics from a BERT tokenizer before and after preprocessing.

Dataset	mean	s.t.d	min	0.25	0.5	0.75	0.9	max
Original	509.32	365.21	23	271	391	640	970	4513
w/ preprocessing	299.80	230.95	12	151	229	366	547	2805

4.2. RoBERTa Model

The proposed system architecture is based on fine-tuning a pre-trained RoBERTa model [22]. Specifically, we utilized the RoBERTa model from the MarIA project [2] [31]. For the classification task, the representation of the final layer's [CLS] token is extracted and passed through a sequence of additional layers: a dropout layer [32], a pooling layer, a Tanh activation, another dropout layer, and a final linear classifier layer. Figure 3 illustrates this architecture.

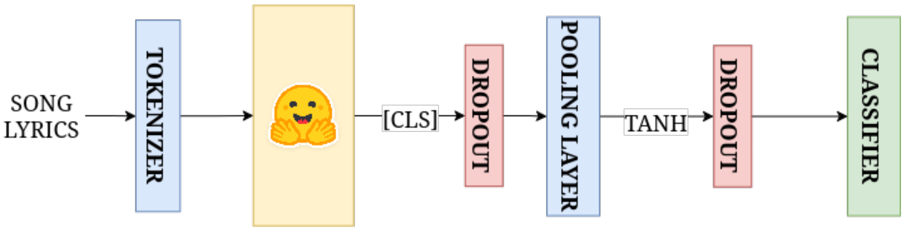


Figure 3: System proposal for both subtasks.

4.3. LLM for Tackling Misogyny Detection

4.3.1. Data Augmentation

As we pointed out in Section 3.2, one of the problems we can anticipate for both sub-tasks is the lack of data for some categories. In order to address this problem, we have decided to employ data augmentation techniques using LLMs to generate new data for the misogynistic lyrics. More specifically, we have explored data augmentation by paraphrasing [33, 34] as well as synthetic data creation [35].

Rather than direct paraphrasing, we explicitly prompt the LLM with a definition of the target fine-grained misogyny category and instruct it to first identify misogynistic verses related to that category. Subsequently, the model generates three distinct paraphrases, with each version adhering to specific rules: the first paraphrases all content while maintaining the core misogynistic theme, the second replaces non-misogynistic verses with new content while only paraphrasing the misogynistic elements, and the third reorganizes and paraphrases the entire song while preserving narrative coherence and the defined misogynistic essence. Prior to generation, the LLM is also tasked with explaining and justifying its intended changes and how the new versions retain the original misogynistic nature, encouraging the model to take intermediate steps before generating any data. Our paraphrase prompt can be found in Appendix 8.1.

On the other hand, for synthetic data creation, our intention was to generate entirely new song lyrics from an existing one, critically ensuring the preservation of the original's misogynistic elements in the new composition. Prior to generating these new lyrics, we instruct the model to extract crucial information: the song's main themes, specific misogynistic verses, its genre, and its overall mood. Subsequently, the model is tasked with creating a new song that retains the same mood and themes but belongs to a different music genre. Our synthetic data generation prompt can be found in Appendix 8.2.

---

[2] PlanTL-GOB-ES/roberta-base-bne

4.3.2. Classification

A primary consideration when designing our classification prompts is that it is not necessary to create a separate prompt for each task. As highlighted in Section 3.2, the second sub-task is dependent on the first. Consequently, successfully addressing the second sub-task enables us to infer the categories for the initial sub-task, which minimizes both employed time and economic costs if we were to use a provider API.

Furthermore, our classification prompts adhere to the conventional zero-shot and one/three-shot schemes for few-shot learning. To enhance the LLM performance, we incorporated CoT reasoning by instructing the model to meticulously analyse song lyrics and formulate its responses through a step-by-step process [28]. Moreover, the model was directed to provide a justification for its classification, which provides us with a plausible explanation for its decision.

Instead of randomly selecting the samples for our in-context learning prompts, we have utilized a model which combines both automatic extraction as well as manual revision. To do so, we first calculated embeddings for all song lyrics with a multilingual embedding model [3] [36] which excels at the Massive Text Embeddings Benchmark (MTEB) [37].

Once embeddings were computed, our goal was to select both representative and ambiguous examples for each class. We achieved this by calculating each song lyric's distance to the centroid of every class. Representative samples were chosen based on their proximity to their own class centroid, providing the LLM with clear examples of what to look for. Conversely, ambiguous samples were picked if they were closer to another class's centroid, exposing the LLM to "difficult" examples.

Given its reliance on semantic similarity, a linguist manually evaluated the top-K representative and ambiguous samples extracted by the algorithm. Thus, the selected samples are guaranteed to represent the phenomena of each of the categories from the second sub-task.

Our 0-shot classification prompt can be found in Appendix **??**.

5. Experimentation Settings

5.1. Data Split

For each task, we created our private training and validation datasets by splitting the original training dataset provided by the organizers, using an 80/20 ratio while stratifying the split to ensure proportionate representation of classes. Subsequently, when applying the data augmentation techniques mentioned on Section 4.3.1, we experimented with various training configurations; for instance, some setups involved replacing all samples of the minority class with synthetic data, whereas others combined both the original misogynistic samples with the synthetic ones. Critically, data augmentation was exclusively applied to samples already present in our private training dataset, ensuring no data leakage into the private validation sets.

5.2. RoBERTa

Our RoBERTa models were fine-tuned following several key recommendations presented in [38]. Specifically, we utilized an AdamW optimizer [39] with a learning rate of $\eta = 2e - 5$; all other AdamW settings were kept at their default values. Moreover, we applied weight decay only to the Transformer encoder layers, but not to the pooling and classification layers. Consistent with these recommendations, the weights of both the pooling and classification layers were initialized with values drawn from a Normal distribution and the selected Transformer configuration (mean = 0, standard deviation = 0.02).

Models were trained with a batch size of 16 for a maximum of 10 epochs. To optimize training duration and prevent overfitting, early stopping was implemented based on the validation loss, and the model checkpoint achieving the lowest validation loss was selected. Input sequences were dynamically padded to the maximum length within each batch to ensure maximum computation efficiency.

---

[3] Alibaba-NLP/gte-multilingual-base

5.3. LLM

All experimentation involving LLMs was conducted using the OpenAI model GPT-4.1-mini [40]. Our selection is based on the fact that this model not only surpasses the GPT-4o model in various benchmarks (e.g., Massive Multitask Language Understanding [41, 42] or Instruction-Following Evaluation [43]), but it also offers its cost-effectiveness.

Regarding more specific settings, we only tweaked the temperature of the model depending on the given task. On the one hand, classification tasks were run with a temperature of 0.5, promoting more consistent and deterministic outputs, yet still allowing the model sufficient flexibility to provide nuanced reasoning, especially for misogyny detection. On the other hand, data augmentation was performed with a temperature of 0.7, thus enabling the model to explore more creative and diverse generations for the resulting data.

6. Results & Discussion

6.1. Quantitative Results

6.1.1. Data Augmentation Quality

Prior to presenting the performance of the system, we evaluated the quality of data generated through augmentation techniques. The results, reflected in Table 3, demonstrate that all the paraphrased data, across all variations, offer a high degree of semantic similarity to the original content, as shown by consistently high BERT-scores. However, the similarity across the three paraphrasing methods is also very close, suggesting that despite specifying rules for diversity, the model did not introduce as much variability as expected.

In contrast, the synthetically generated data stands out for its lower cosine similarity, indicating greater lexical and structural diversity. Despite this, its BERT-score remains relatively high (0.76), implying that the generated content still retains much of the original meaning.

Table 3
Cosine similarity and BERT-score values for the proposed data augmentation techniques.

Data augmentation technique	Cosine Similarity ↑	BERT-score ↑
Paraphrase 1	0.87	0.79
Paraphrase 2	0.85	0.78
Paraphrase 3	0.85	0.78
Synthetic data	0.68	0.76

6.1.2. Misogyny Speech Detection Results

Table 4 displays that our RoBERTa model trained on the original dataset consistently achieves the best overall performance across most metrics. When the original data is replaced with paraphrased versions (Par. {1,2,3}), a performance drop is observed in all cases, with the third paraphrasing variant showing the most pronounced decline. A similar trend is evident when training solely on synthetic data (Synthetic), which also underperforms compared to the original configuration. Furthermore, combining the original dataset with either paraphrased (Original + Par. {1,2,3}) or synthetic data (Original + Synthethic) does not yield improvements over the baseline. This suggests that the augmented samples may introduce noise or redundancy rather than helpful variability, hindering model performance.

Table 5 presents the results for the GPT-4.1-mini model across different prompt configurations. As anticipated, increasing the number of examples in the prompt leads the LLM model to yield results that more closely resemble those obtained by our RoBERTa-based approach, indicating the benefit of in-context learning. Note that the LLM results on classification do not incorporate the AUROC metric, since it relies on class probabilities in order to calculate its value.

Table 4

Results for our RoBERTa model depending on the training dataset configuration used. Bolded results highlight the best model for each metric.

Dataset Configuration	AUROC ↑	MCC ↑	F1 ↑	Pr ↑	Rc ↑
Original	**0.87**	**0.59**	**0.79**	0.79	**0.80**
Par. 1	0.81	0.42	0.71	0.71	0.71
Par. 2	0.81	0.43	0.70	0.77	0.67
Par. 3	0.75	0.04	0.41	0.38	0.35
Original + Par. 1	0.87	0.55	0.77	0.77	0.78
Original + Par. 2	0.86	0.55	0.77	0.79	0.76
Original + Par. 3	0.87	0.53	0.76	0.77	0.76
Synthetic	0.85	0.51	0.72	**0.82**	0.70
Original + Synthethic	0.83	0.53	0.76	0.80	0.74

Table 5

Results depending on the prompt used for GPT-4.1-mini. Bolded results highlight the best model for each metric.

Prompt	MCC ↑	F1 ↑	Pr ↑	Rc ↑
0-shot	0.52	0.69	0.80	0.73
1-shot	0.56	0.74	0.80	0.76
3-shot	**0.61**	**0.78**	**0.82**	**0.80**

Figure 4 displays the confusion matrices for the evaluated models. The RoBERTa model demonstrates consistent performance across both classes, exhibiting a stronger tendency to correctly identify non-misogynistic samples. In contrast, GPT-4.1-mini struggles with detecting misogynistic content, especially in the 0-shot settings, where its predictions are heavily skewed towards the non-misogynistic class. Although performance improves slightly with more examples, the LLM still does not match RoBERTa's balance between the two categories.

For the official evaluation performed on the test set, our submissions were sent from the predictions given by the default dataset configuration and RoBERTa as well as all the proposed prompts using GPT-4.1-mini. Table 6 presents our official test results. Our RoBERTa model secured fifth place in the overall competition, with its performance on the test set aligning well with observations from the validation set evaluation. While the LLM classification performance remains lower than RoBERTa, we observed improvements in both the 0-shot and 1-shot scenarios in the test set compared to validation. However, the 3-shot results were consistent with those from the validation set.

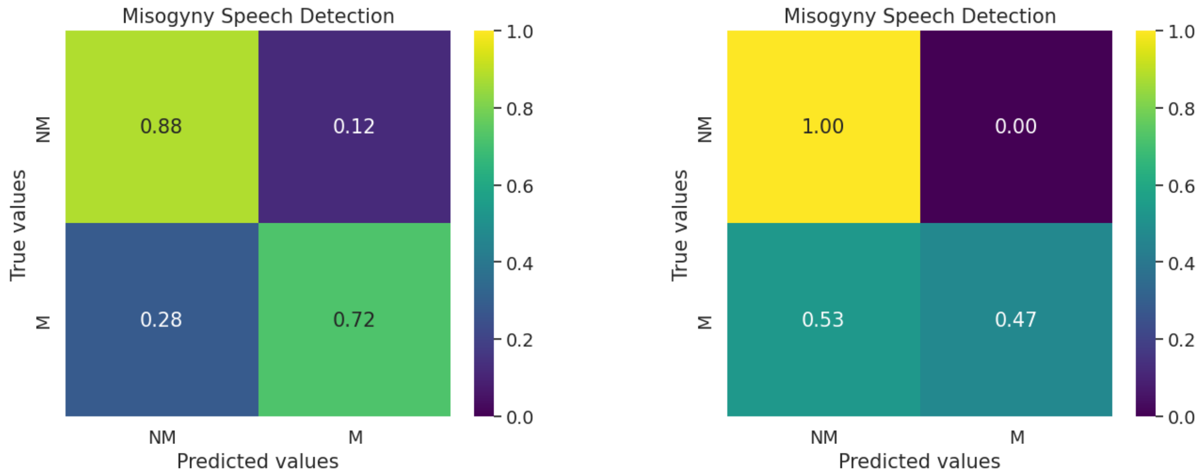
Table 6

Official test results for the Misogyny Hate Speech subtask.

Model	Ranking	F1 ↑
RoBERTa (ours)	5	0.8039
0-shot (ours)	39	0.7479
1-shot (ours)	20	0.7947
3-shot (ours)	17	0.7987
Winner	1	0.8811

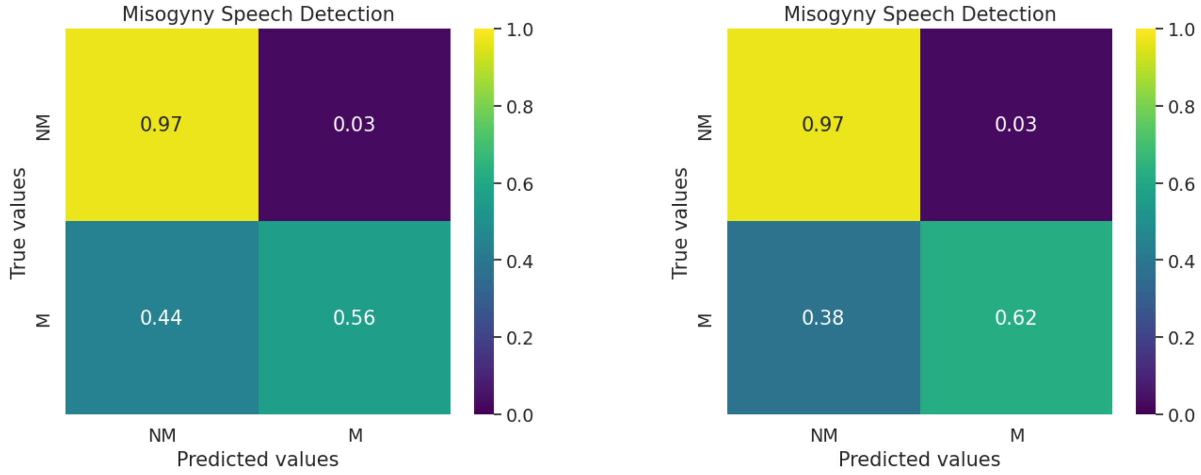
6.1.3. Fine-grained Misogyny Speech Detection

Table 7 presents the results of the different dataset techniques utilized compared to the original dataset for the RoBERTa model. As can be seen, the results are very similar to the ones we observed earlier.



(a) RoBERTa model with the original dataset.

(b) 0-shot with GPT-4.1-mini.



(c) 1-shot with GPT-4.1-mini.

(d) 3-shot with GPT-4.1-mini.

Figure 4: Confusion matrix normalized by true values for the Misogyny Speech Detection sub-task.

In other words, replacing the original data by the generated from the LLM (Par. {1,2,3} and Synthetic) offers a worse performance in comparison with the original baseline. However, a tiny performance improvement can be observed once we combine the generated data from the LLM alongside the original data.

Table 7

Results for our RoBERTa-based approach on the Fine-grained Misogyny Speech Detection task depending on the training configuration used. Bolded results highlight the best model for each metric.

Dataset	AUROC ↑	MCC ↑	F1 ↑	Pr ↑	Rc ↑
Original	**0.77**	0.47	0.37	0.34	0.41
Par. 1	0.65	0.34	0.31	0.33	0.34
Par. 2	0.64	0.32	0.30	0.33	0.33
Par. 3	0.60	0.28	0.31	0.28	0.34
Original + Par. 1	0.74	**0.48**	0.37	0.34	0.41
Original + Par. 2	0.76	0.47	0.38	0.38	**0.42**
Original + Par. 3	0.76	0.45	0.38	0.39	0.41
Synthetic data	0.50	0.05	0.15	0.11	0.25
Original + Synthethic	**0.77**	0.47	**0.39**	**0.42**	**0.42**

Table 8 presents the results for the Fine-grained Misogyny Speech Detection using GPT-4.1-mini with different prompting strategies. As shown, the LLM consistently outperforms our RoBERTa system-based approach for all prompts, including 0-shot.

Table 8
Results for the Fine-grained Misogyny Speech Detection task given different prompts.

Prompt	MCC	F1 ↑	Pr ↑	Rc ↑
0-shot	0.47	0.47	0.68	0.46
1-shot	0.50	0.51	0.64	0.49
3-shot	**0.57**	**0.57**	**0.75**	**0.73**

To further analyse the results, we examined the system outputs using Figure 5. For the RoBERTa model with the original dataset (Figure 5a), high recall is observed for NR and S, which suggests that the model is overfitting on these categories since they are the most predominant on this task. However, the model significantly struggles with hate and violence, showing a very low recall of these crucial categories, frequently misclassifying them as NR or S.

The LLM model's confusion matrices (Figures 5b, 5c, and 5d) show a progressive improvement with more examples. While initially biased towards predicting NR (especially in 0-shot), recall for S consistently improves across all scenarios. H also sees a significant recall improvement from 0-shot to 3-shot. However, the V category remains challenging for the LLM, showing only marginal improvements in recall across settings. We will analyse specific outputs for the V category in the subsequent section to better understand these limitations.

For the official test set, we sent the predictions from the RoBERTa model trained with the original training dataset configuration as well as the LLM predictions using the different prompts. Table 9 reflects our results in the official test set as well as a comparison with the winner of the subtask. In addition, our test results prove that the LLM proposal offers better results in the given task. Moreover, our best submission is the 3-shot prompt, which has achieved a third position in the overall ranking.

Table 9
Official test results for the Fine-grained Misogyny Hate Speech subtask.

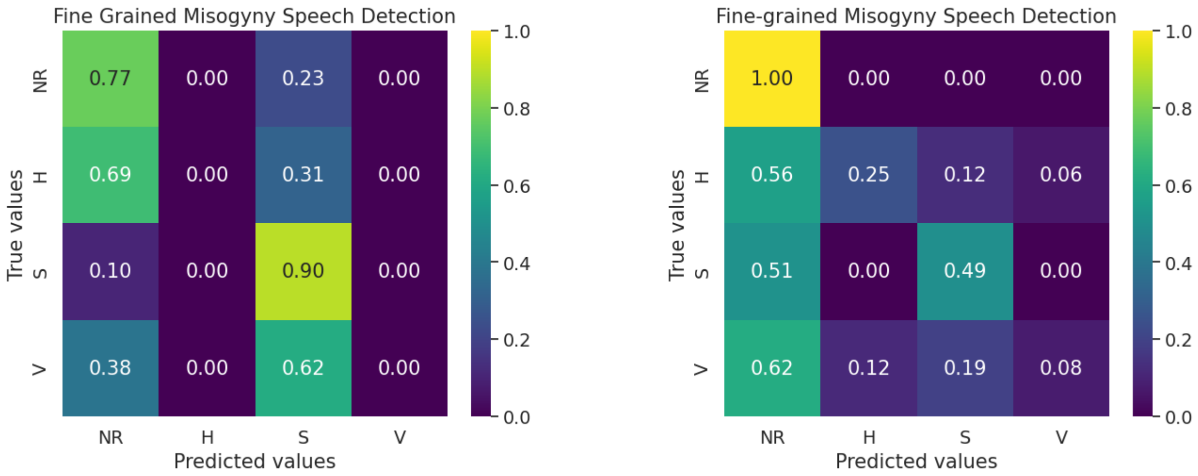
Model	Ranking	F1 ↑
RoBERTa (ours)	31	0.3816
0-shot (ours)	17	0.4556
1-shot (ours)	8	0.5254
3-shot (ours)	3	0.5459
Winner	1	0.5895

6.2. Qualitative Results

6.2.1. Data Augmentation Quality Analysis

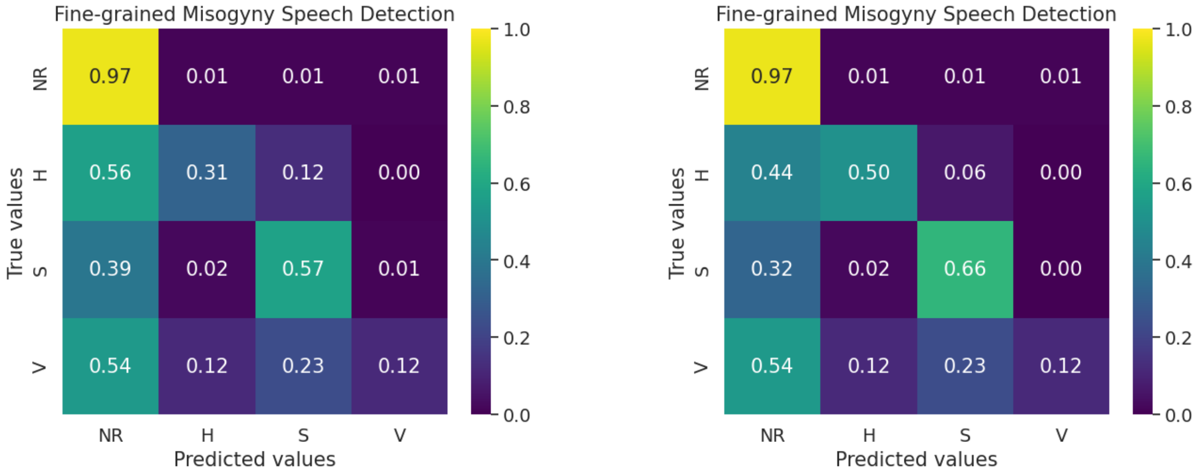
The concise results explained in the last Section 5.1, provide a rigorous context for a qualitative analysis that allows us to explore the GPT-4.1-mini data-augmentation behaviour. In this subsection, we explore the rationale of the LLM behind the generation of paraphrased samples. Table 10 shows an example of a paraphrased song.

In general, taking into account the conditions of the paraphrasing prompt, we observe three main issues. Firstly, there is a lack of creativity when the LLM tries linguistic paraphrasing strategies. The three versions of the paraphrase show poor writing innovation reflected in verses that are almost equal between versions. For instance, "No matter how far you are", "even if you are far", "even if you're far away". The LLM only adds some synonyms or changes the verb tense keeping some of the verses almost identical to the original song. Secondly, we noticed that the length of the paraphrases is significantly shorter than the original (especially in the third paraphrase) as it can be observed in Table 10. Lastly, after exploring the generated samples, we noticed that the LLM does not apply the condition of the



(a) RoBERTa model with the original dataset.

(b) 0-shot with GPT-4.1-mini.



(c) 1-shot with GPT-4.1-mini.

(d) 3-shot with GPT-4.1-mini.

Figure 5: Confusion matrix normalized by true values for the Fine-Grained Misogyny Speech Detection sub-task.

third paraphrase in the prompt: "[...] paraphrase all misogynistic and non-misogynistic verses of the song, but modify the structure of the song as well as the order of the verses [...]", as it can be seen from the third example in the table, the order and structure of verses are kept the same as in the original.

In conclusion, after our exploration of the data augmentation results, we can infer that creative writing is an outstanding challenge in generative AI. This is an ongoing discussed subject ([44], [45], [46]) in which the limits of the creative thinking in AI are still unclear.

6.2.2. Misogyny Speech Classification Analysis

In this subsection, we further present examples of the responses produced by the GPT-4.1-mini when detecting the categories from the second sub-task (NR, S, H, V) to the different prompts: 0-shot, 1-shot and 3-shot. With this aim, we establish a comparison among the three different prompting strategies and explore the performance results in detail. The following three tables present three significant examples of the response of the model in different categories. Some of them include interesting information about the detailed reasoning of GPT-4.1 when making a classification decision. For instance, we can observe the presence of semantic arguments and linguistic expressions to justify the LLM choice and how its interpretation of the semantic context of the song affects its final decision.

Table 11 presents the responses to a song categorized as "Sexualization". As we mentioned in Section 5.1, this is one of the predominant categories within the dataset, but even though the LLM performance

Table 10

Comparison of paraphrased versions of ID T2_TRAIN_0193 (S)

T2_TRAIN_0342

Original:

<Estoy tan lejos de ti Y a pesar de la enorme distancia Te siento juntito a mí; Corazón, corazón; alma con alma Y siento en mi ser tus besos No importa que estés tan lejos Estoy pensando en tu amor Y a lo loco platico contigo Te cuento de mi dolor Y aunque me hagas feliz No te lo digo Y vuelvo a sentir tus besos No importa que estés tan lejos El cielo empieza a clarear Y mis ojos se llenan de sueño Contigo voy a soñar Porque, quieran o no Yo soy tu dueño Y siempre tendré tus besos No importa que estés tan lejos Estoy pensando en tu amor.>

Paraphrase 1:

<Me siento lejos de tu piel Aunque estemos distantes, Te imagino bien cerca de mí; Alma con alma, latidos juntos, Y en mi cuerpo guardo tus caricias, No importa que estés tan lejos. Mi mente se llena de tu amor, Hablo contigo en mi delirio, Te comparto mi tristeza, Aunque no lo diga, tú me completas, Y vuelvo a sentir tus labios, No importa que estés tan lejos La madrugada comienza a salir, Los párpados se ciegan de cansancio, Contigo quiero soñar, Porque, guste o no, soy quien te posee, Y siempre tendré tus besos, No importa que estés tan lejos.>

Paraphrase 2:

<Estoy lejos de tu figura Pienso en ti sin medida, Siento tu calor en la distancia, Corazón junto al mío, Tus caricias me persiguen, Aunque estés lejos de aquí. Me absorto en las llamas de tu querer, Hablo contigo en mis sueños, Te revelo mi tristeza, Aunque no lo confiese, Y tus labios son míos, No importa que estés lejos. Cuando el alba empieza a brillar, Mis ojos se cierran de fatiga, Quiero perderme en tus sueños, Porque, quieras o no, Soy el dueño de tu cuerpo, Y siempre poseeré tus besos, Aunque estés lejos.>

Paraphrase 3:

<Soy quien controla tu cuerpo Y siempre tendré tus besos, No importa donde estés. Aunque la distancia crezca, Te siento muy cerca de mí, En mi alma estás presente, Y tus caricias me envuelven. Hablo contigo en silencio, Revelo mi pena oculta, Y aunque no te lo diga, Mi deseo por ti no muere. Al amanecer me rindo al sueño, Cierro los ojos imaginando, Que tus labios son míos, Aunque estés muy lejos.>

shows high results, there are some cases where it struggles to detect "Sexualization". As we observe, this song is classified as "Non-Related" in 0-shot and 1-shot responses. The LLM gives different arguments to support its classification decision, but some of them are inconsistent: "using affectionate and flattering language (for example: 'you're perfect', 'doll, princess', 'I love all of you'), without dehumanizing or objectifying" (0-shot) and "barbie, doll, princess physical descriptions are expressed [...] without any degrading intent or objectification"(1-shot). In both responses, linguistic cues carry explicit meanings of women body objectification that are interpreted as mutual sexual desire. This is reformulated in the 3-shot response in which the model improves its performance considering the association of women with figures such as Barbie or a doll as clearly objectifying and subordinating women.

The second table (Table 12) shows a comparison of the LLM responses to a song categorized as "Violence" by the true label. This particular example demonstrates how the model struggles to detect instances of violent actions against women even when they are explicitly expressed through linguistic cues. As we observe, the model misclassifies this category as "Sexualization" and tries to justify the presence of some linguistic expressions that carry significant semantic weight such as "I could be your rapist", "I'll take you to the limit" or "I'll make you move to extremes". These expressions are detected by the LLM as sexual objectification or sexual subordination but are understood as sexual and dominance play within the context of "Sexualization". In this sense, the responses of 0-shot and 1-shot tend to misunderstand the violence expressed by the song masked under sexualization and domination play. In addition, in 3-shot the model mitigates the violent connotation of expressions such as "I could be your rapist" and supports its decision mentioning that these linguistic instances carry an implicit meaning of

Table 11

Comparison of reasoning responses generated by different prompts for category S

T2_TRAIN_0342

0-shot:

<The lyrics of the song "Es Un Secreto" do not meet the general definition of misogyny [...] Nor does it perpetuate gender stereotypes that promote subordination [...]. The content is primarily romantic [...] using affectionate and flattering language (for example: "you're perfect," "**doll**, princess," "I love all of you"), **without dehumanizing or objectifying women** in a degrading way. [...] **The portrayal of sexuality is not used as a tool for domination** or degradation, but rather as an expression of desire and connection between two people. Consequently, the correct classification is 'Not Related' (NR).>

1-shot:

<The lyrics of "Es Un Secreto" do not meet the established definition of misogyny, [...] **Nor do they perpetuate stereotypes of female subordination or objectify the body** or its parts in a degrading manner [...]. The use of metaphors such as "**Barbie**, **doll**, princess" and physical descriptions ("your face, your body, your eyes, also your skin") is expressed in a **tone of admiration and appreciation, without degrading intent or objectification** [...] the song conveys an ideal of connection between two parties where desire is shared and consensual. Therefore, the song should be classified as 'Not Related' (NR)[...].>

3-shot:

<The song "Es Un Secreto" **is classified as 'Sexualization'(S)** because, [...] **it does perpetuate the objectification and subordination of women** through sexualization. The lyrics **repeatedly highlight female physical attributes** [...] and use terms like "**Barbie**, **doll**, princess," **which objectify women by comparing them to idealized and superficial figures**. Additionally, there is **an underlying intention of domination** [...], and the invitation to "dance sensually" and "please," which **suggest a dynamic where the woman is primarily seen as an object of sexual pleasure**. [...] it fits the definition of misogyny in terms of promoting the subordination and objectification of the female body **with an implicit intent of sexual domination**, clearly placing it within the category of Sexualization.>

sexual subordination of women. In conclusion, this is one of the examples that illustrates how "Violence" against women goes undetected by the model.

Finally, the third table presents a comparison of the model response to a song in the category "Hate". "Hate" is one of the categories from sub-task two that improves with 3-shot while it presents some performance limitations in 0-shot and 1-shot according to the confusion matrix presented in Figure 5 (5b, 5c, 5d). The following example illustrates this scenario in which the song is classified as "Non-Related" in 0-shot and 1-shot. According to the model, in both responses, the song "contains neither explicit disdain nor violent intentions against women". In addition, in 0-shot, the model supports the idea that "although the narrator expresses a certain degree of contempt or resentment due to having been abandoned, the language used is neither offensive nor discriminatory toward women". Also, in 1-shot, it states a similar reason: "Pejorative language is used to describe the man with whom the woman was involved, such as the expression ('that bastard')[...]the underlying intention seems to align more with a narrative centered on betrayal and personal empowerment." There is a significant improvement in 3-shot since the model recognizes linguistic cues that carry explicit expressions of misogyny and hate speech against women such as "Consider yourself dead; your time is up." Also, in 3-shot, the model interprets the expression "that bastard" as one of the expressions conveying "disdain and resentment toward the woman."

From a qualitative perspective, we consider two key aspects from the examples analyzed in this section. On the one hand, there is a clear improvement in the classification performance of the model. As previously discussed, the category of "Sexualization" is the most accurately classified, for the reasons outlined earlier. However, in some instances, the model overlooks certain linguistic cues that may be crucial for detecting the appropriate category. The categories of "Hate" and "Violence" exhibit notable

Table 12
Comparison of reasoning responses generated by different prompts for category V

T2_TRAIN_0327

0-shot:
<The song contains language that **suggests sexual acts and domination games**, specifically in verses like “**I’ll be your butler and you’ll play the role of Lady Fine**” and “**Or I could be your rapist**[...], which imply a sexual relationship based on domination[...]. There are also explicit references [...]**which objectify the female body with a clearly sexual intent**. Although the lyrics use figurative language and **a framework of a “seduction game,” the depiction of scenes in which the woman is objectified and placed in a subordinate role during sexual acts fits the definition of misogyny** through sexualization. **No direct expressions of physical or psychological aggression** (Violence category) or explicit hate speech (Hate category) are identified. Therefore, the appropriate classification is Sexualization(S).>

1-shot:
<The song "Juego de Seducción" is classified as Sexualization (S) [...] the woman is portrayed in a **passive or subordinate role** within an **implicitly dominant context**. [...]especially the line “**Or I could be your rapist**, imagination” **suggest a power dynamic that implies an unequal relationship** in which **the woman is positioned as an object within a sexual game of domination and submission**. [...] “rapist” [...] directly **invokes a violent sexual act** [...]. Furthermore, the chorus [...] reinforce the dominant and controlling intent **from a sexual angle** that instrumentalizes female sexuality. **There is no direct threat of physical violence** or clear expression of hate, **so it should not be classified under Violence (V)** or Hate (H). [...]. For these reasons, it clearly fits under the category Sexualization (S).>

3-shot:
<The song "Juego de Seducción" meets the definition of misogyny [...] the lyrics suggest roleplay [...] and **explicitly mention the possibility of being “your rapist”, which implies violent intent and submission in a sexual context**. Although the **violence is not directly expressed** as a real threat or harm, **the reference to rape within a game scenario sexualizes the woman as a passive, subordinate object, reinforcing subordination** [...] “I’ll take you to the edge” and “your very skin” are **expressions that objectify the female body in a context of sexual domination**. Therefore, the misogyny present is mainly expressed through sexualization, with an implicit component of domination and submission[...]>

classification limitations and reveal that the arguments developed by the model are significantly more biased in these cases compared to the “Sexualization” category. Nevertheless, while the classification of “Hate” improves considerably in responses using 3-shot prompting, the model still struggles to accurately classify songs containing semantic content related to violence against women. On the other hand, after a detailed examination of the generated responses, the misclassification of “Violence” as “Sexualization” or “Non-Related” may be due to the presence of more linguistically complex structures that convey implicit meanings, which are far more subtle than the explicit language found in songs classified as "Sexualization". In addition, we could suggest that the model may exhibit greater bias when interpreting this type of content as a consequence of the perpetuation of gender biases in AI. The presence of misogyny is a one of the most complex issues to be detected since it requires a nuanced understanding of language and context [47]. The reason behind the presence of gender biases in LLMs is that the training data used to develop a language model is made of human-produced texts containing different types of sociocultural biases held by humans [48]. In this sense, further work needs to be done to constitute a more comprehensive research to evaluate debiasing methods to mitigate the presence of gender and other social bias in AI.

Table 13
Comparison of reasoning responses generated by different prompts for category H

T2_TRAIN_0452

0-shot:
<[...] Colloquial language and expressions reflecting the narrator’s **discontent and perceived superiority over** another person (presumably **a woman**, given the reference to a past romantic relationship) are evident. However, the lyrics **do not contain hate speech, explicit contempt, or any violent intent** directed toward the woman or toward women as a group. [...] Although the narrator expresses a certain degree of disdain or rejection due to being abandoned, **the language used** [...] **does not employ hostile terms that would suggest hate speech.** Therefore, the song does not meet the strict definition of misogyny provided and should thus be classified in the "Not Related" (NR) category.>

1-shot:
<The lyrics of the song do not meet the provided definition of misogyny. The text refers to a woman who allegedly was with another man, expressing a sense of superiority and triumph on the part of the speaker (“I’m the boss”), and using derogatory terms to refer to the man involved with the woman (“that bastard”). However, **the lyrics do not contain hate speech** [...] there are no expressions of violence or threats directed at her. Furthermore, **the underlying intention appears to be more about a narrative of betrayal and personal empowerment**, **rather than misogynistic hostility** [...] should be labeled as “Not Related” (NR).>

3-shot:
<**The song meets the general definition of misogyny by containing expressions of contempt and hostility toward a specific woman**. The language used-such as **“bastard” and phrases that convey scorn and resentment toward the woman** [...]—**reflects a discourse of hatred and disdain directed at her**. Additionally, the phrase **“Consider yourself dead, your time is up” can be interpreted as an expression of absolute rejection** [...] it **does not include an explicit threat** of physical violence, **it implies a desire to erase or eliminate her presence** [...] **the most fitting category is Hate (H)**, as the lyrics reflect offensive language, disdain, and resentment toward a woman.>

7. Conclusions and Future Work

This work has addressed the detection of misogyny—an already well-explored topic in NLP—within the under-researched yet culturally significant domain of song lyrics. By participating in the MiSonGyny 2025 shared task, we contribute to expanding misogyny detection into new discursive contexts.

Our experiments with a RoBERTa-based system confirm that fine-tuned transformer architectures remain effective for this task, achieving fifth place in the general misogyny detection subtask. However, their performance suffers under data imbalance, particularly in the fine-grained classification scenario, where overfitting is more pronounced, achieving only 31st position on the test ranking.

In contrast, our exploration of LLMs investigated the impact of in-context learning and chain-of-thought prompting. Despite limited data, LLMs demonstrated strong performance in the fine-grained subtask, securing third place and outperforming BERT-like models. Beyond raw performance, we analysed the LLM’s reasoning capabilities, noting that while it effectively identified explicit categories like Sexualization (S), it struggled with more implicit ones like Violence (V) and Hostility (H). These difficulties may stem from the subtlety of these expressions and the model’s inability to justify its predictions with coherent rationale—issues clearly reflected in the confusion matrix.

Furthermore, we examined the use of LLMs for data augmentation through paraphrasing and synthetic sample generation. However, the results suggest limited variability in the augmented data, potentially due to constraints imposed by safety filters or a lack of creativity in LLM-generated text [49, 17].

Building on our LLM experimentation, we plan to extend our work beyond closed models, utilizing open LLMs like the Gemma [50] or Llama [25] series. This will not only enhance reproducibility for the scientific community but also offer valuable insights into their comparative performance.

On the other hand, although the following work addresses the detection of misogyny in song lyrics, we believe there is future research to be conducted regarding the detection of stereotypes [51, 52] in song lyrics.

In conclusion, our study presents a multifaceted evaluation of misogyny detection in lyrics, comparing traditional fine-tuned models and LLMs, and highlighting their respective strengths and limitations. We provide both quantitative metrics and qualitative insights, offering new perspectives on LLM reasoning and their potential for future work in data augmentation and interpretability.

Acknowledgments

This work has been developed under the framework of the following projects respectively: MARTINI (Grant PCI2022-135008-2) and FairTransNLP-Stereotypes (Grant PID2021-124361OB-C31) both funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

Declaration on Generative AI

During the preparation of this work, the authors used used ChatGPT and DeepL in order to: Text Translation, Paraphrase and reword and Improve Writing style. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Harsiwi, R. Fatimah, F. Alfarisy, Gendered Hate: How Affective Language Shapes Hate Speech Against Women, International Journal of Linguistics, Literature and Translation 7 (2024) 17–21.

[2] J. T. Nockleby, Hate speech, Encyclopedia of the American constitution 3 (2000) 1277–1279.

[3] L. G. Moreno-Sandoval, A. Pomares-Quimbaya, S. A. Barbosa-Sierra, L. M. Pantoja-Rojas, Detection of Hate Speech, Racism and Misogyny in Digital Social Networks: Colombian Case Study, Big Data and Cognitive Computing 8 (2024) 113.

[4] A. Cascione, A. Cerulli, M. M. Manerba, L. Passaro, Women's Professions and Targeted Misogyny Online, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 182–189.

[5] S. Davis, Pop Lyrics: A Mirror and a Molder Of Society, Et cetera 42 (1985) 167.

[6] L. Betti, C. Abrate, A. Kaltenbrunner, Large Scale Analysis of Gender Bias and Sexism in Song Lyrics, EPJ Data Science 12 (2023) 10.

[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] F. Rodríguez-Sánchez, J. Carrillo-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.

[10] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: Sexism Identification in Social Networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[11] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023-Learning with Disagreement for Sexism Identification and Characterization (Extended Overview)., CLEF (Working Notes) (2023) 813–854.

[12] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024–Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview), in: Conference and Labs of the Evaluation Forum, 2024.

[13] W. S. Schmeisser-Nieto, P. Pastells Vilà, S. Frenda, A. Ariza-Casabona, P. Rosso, M. Taulé Delor, et al., Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish-Learning with Disagreement, Procesamiento del Lenguaje Natural 73 (2024) 323–333.

[14] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), volume 2150 of *CEUR Proc.*, Seville, Spain, 2018, pp. 214–228.

[15] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.

[16] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, Advances in neural information processing systems 33 (2020) 2611–2624.

[17] D. Chen, A. Satish, R. Khanbayov, C. Schuster, G. Groh, Tuning Into Bias: A Computational Study of Gender Bias in Song Lyrics, in: Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), 2025, pp. 117–129.

[18] L. Casanovas-Buliart, P. Alvarez-Cueva, C. C. and, Evolution over 62 years: An Analysis of Sexism in the lyrics of the most-listened-to songs in Spain, Cogent Arts & Humanities 11 (2024) 2436723. URL: https://doi.org/10.1080/23311983.2024.2436723. doi:`10.1080/23311983.2024.2436723`. `arXiv:https://doi.org/10.1080/23311983.2024.2436723`.

[19] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, 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.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Advances in neural information processing systems 30 (2017).

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL: https://arxiv.org/abs/1907.11692. `arXiv:1907.11692`.

[23] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, arXiv preprint arXiv:2412.13663 (2024).

[24] O. Team, GPT-4 Technical Report, 2024. URL: https://arxiv.org/abs/2303.08774. `arXiv:2303.08774`.

[25] M. AI, The Llama 3 Herd of Models, 2024. URL: https://arxiv.org/abs/2407.21783. `arXiv:2407.21783`.

[26] P. Törnberg, Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages, Social Science Computer Review 0 (0) 08944393241286471. doi:`10.1177/08944393241286471`.

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, Advances in neural information

processing systems 33 (2020) 1877–1901.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-Of-Thought Prompting Elicits Reasoning In Large Language Models, Advances in neural information processing systems 35 (2022) 24824–24837.

[29] A. P. Bradley, The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms, Pattern Recognition 30 (1997) 1145–1159. doi:`https://doi.org/10.1016/S0031-3203(96)00142-2`.

[30] D. Chicco, G. Jurman, The Advantages Of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation, BMC Genomics 21 (2020) 6. doi:`10.1186/s12864-019-6413-7`.

[31] A. Gutiérrez Fandiño, J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento del Lenguaje Natural 68 (2022).

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way To Prevent Neural Networks from Overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[33] E. Okur, S. Sahay, L. Nachman, Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4114–4125. URL: https://aclanthology.org/2022.lrec-1.437/.

[34] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, F. Zeng, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, X. Li, AugGPT: Leveraging ChatGPT for Text Data Augmentation, IEEE Transactions on Big Data 11 (2025) 907–918. doi:`10.1109/TBDATA.2025.3536934`.

[35] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 10443–10461.

[36] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, et al., mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2024, pp. 1393–1412.

[37] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive Text Embedding Benchmark, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2014–2037. doi:`10.18653/v1/2023.eacl-main.148`.

[38] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, Revisiting Few-sample BERT Fine-tuning, in: International Conference on Learning Representations, 2021.

[39] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations (ICLR), 2019.

[40] OpenAI, Introducing GPT-4.1 in the API, https://openai.com/index/gpt-4-1/, 2025. Accessed on 19-05-2025.

[41] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring Massive Multitask Language Understanding, Proceedings of the International Conference on Learning Representations (ICLR) (2021).

[42] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI With Shared Human Values, Proceedings of the International Conference on Learning Representations (ICLR) (2021).

[43] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, L. Hou, Instruction-Following Evaluation for Large Language Models, 2023. URL: https://arxiv.org/abs/2311.07911. `arXiv:2311.07911`.

[44] Art or Artifice? Large Language Models and the False Promise of Creativity, author=Chakrabarty,

Tuhin and Laban, Philippe and Agarwal, Divyansh and Muresan, Smaranda and Wu, Chien-Sheng, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–34.
[45] Q. Wan, S. Hu, Y. Zhang, P. Wang, B. Wen, Z. Lu, "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–26.
[46] A. Bellemare-Pepin, F. Lespinasse, P. Thölke, Y. Harel, K. Mathewson, J. A. Olson, Y. Bengio, K. Jerbi, Divergent Creativity in Humans and Large Language Models, arXiv preprint arXiv:2405.13012 (2024).
[47] D. Battista, J. C. Molano, et al., How ai bots have reinforced gender bias in hate speech, Ex aequo (2023) 53–68.
[48] A. Caliskan, P. P. Ajay, T. Charlesworth, R. Wolfe, M. R. Banaji, Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 156–170.
[49] M. Ismayilzada, C. Stevenson, L. van der Plas, Evaluating Creative Short Story Generation in Humans and Large Language Models, arXiv preprint arXiv:2411.02316 (2024).
[50] Team, Gemma and Kamath, Aishwarya and Ferret, Johan and Pathak, Shreya and Vieillard, Nino and Merhej, Ramona and Perrin, Sarah and Matejovicova, Tatiana and Ramé, Alexandre and Rivière, Morgane and others, Gemma 3 Technical Report, arXiv preprint arXiv:2503.19786 (2025).
[51] E. U. Alacreu, P. Rosso, Identification of Racial and Sexist Stereotypes in Spanish: A Learning with Disagreements Approach, Procesamiento del Lenguaje Natural 74 (2025) 15–31.
[52] A. T. Cignarella, A. Giachanou, E. Lefever, Stereotype detection in natural language processing, arXiv preprint arXiv:2505.17642 (2025).

letras de canciones que no pertenecen a ninguna de las categorías anteriores (es decir, no cumplen la definición de misoginia). </descripcion>

8. Data augmentation

8.1. Paraphrasing prompt

OBJECTIVE:Generate three paraphrases of the lyrics of the following song, which contains misogynistic elements and/or hate speech, specifically of the type {TYPE_MISOGYNY}. The output must be structured in XML format.
DEFINITION OF MISOGYNY: We speak of a song's lyrics as containing misogyny when they contain hate speech or disdain directed at women. Lyrics that promote the subordination or objectification of women also fall into this category.
DEFINITION OF {MISOGYNY_TYPE}:
{MISOGYNY_TYPE_DEFINITION}.
ADDITIONAL RULES:

- Do not remove misogynistic and/or problematic elements from any of the new versions.
- You do not need to respect the rhyme of the verses, but you do need to keep the central theme as well as the misogyny of {TYPE_MISOGYNY} in the song.
- For the first paraphrase, paraphrase all misogynistic and non-misogynistic content, but keep the hate speech nature of {TYPE_MISOGYNY} as per the definition provided.
- For the second paraphrase, it paraphrases all the misogynistic verses of the song, while the non-misogynistic verses are replaced by entirely new ones in line with the original subject matter of the song as well as its message. The paraphrases of the misogynistic elements should be in accordance with the definition of {TYPE_MISOGYNY} provided.
- For the third paraphrase, paraphrase all misogynistic and non-misogynistic verses of the song, but modify the structure of the song as well as the order of the verses, always

maintaining a narrative coherence that resembles the original lyrics but conforms to the definition of {TYPE_MISOGYNY} provided.

- Before paraphrasing, think about the changes you are going to make and explain them briefly. Briefly justify why your new version still retains the original essence as well as why it is misogynistic according to the definition of {TYPE_MISOGYNY} we have provided.

OUTPUT FORMAT: Your response must strictly follow the following XML format:
<result>
<analysis>
<identification_misogynistic_elements>
<!– Description of the misogynistic elements identified in the original song –>
</identification_misogynistic_elements>
</analysis>
<paraphrasing>
<version number='1'>
<explanation>
<!– Explanation of the changes made and justification of why it retains the misogynistic essence –>
</explanation>
<letter>
<!– Full lyrics of the first paraphrased version –>
</letter>
</version>
<version number='2'>
<explanation>
<!– Explanation of the changes made and justification of why it retains the misogynistic essence –>
</explanation>
<letter>
<!– Full lyrics of the second paraphrased version –>
</letter>
</version>
<version number='3'>
<explanation>
<!– Explanation of the changes made and justification of why it retains the misogynistic essence –>
</explanation>
<letter>
<!– Full lyrics of the third paraphrased version –> </letter>
</version>
</paraphrasing>
</result>
Here are the lyrics to the song: lyrics

8.2. Synthetic data generation prompt

OBJECTIVE: Generate a new song lyric from the lyrics of the following song, which contains misogynistic elements and/or hate speech, specifically of the type {MISOGYNY_TYPE}.
DEFINITION OF MISOGYNY: We speak of a song's lyrics as containing misogyny when they contain hate speech or disdain directed at women. Lyrics that promote the subordination or

objectification of women also fall into this category.
DEFINITION OF {MISOGYNY_TYPE}: {MISOGYNY_TYPE_DEFINITION}.
RULES:

- Preserving misogyny: It is imperative that the song contains similar misogynistic elements in its new versions. The specific type of misogyny {MISOGYNY_TYPE} and the essence of the message must remain clearly recognisable.
- Maintain Original Themes: The newly generated song should respect much of the original themes proposed in the song. For example, if the song is about heartbreak, the new song should focus on this theme, while if it is about sexual acts, the new version should talk about this as well.
- Rhymes: The new version does not need to respect the rhymes between verses, but it does need to maintain misogynistic elements in its final version.
- Mood: Use the feeling/mood associated with the original song to generate the new song. If the song emanates a feeling of sadness, the new song should maintain this essence, while if it emanates a feeling of revenge or hatred, the new song should reflect these feelings.
- Rhetorical figures of speech: Introduce at least 2 or 3 metaphors or similes in the song. However, these figures of speech should NOT affect the explicitness of the misogyny, if any.
- Length: The resulting song should be approximately the length of the original song in terms of the number of verses.
- DON'T JUST PARAPHRASE OR COPY THE ORIGINAL SONG: Generate a new song from the one provided, BUT don't mimic the original song in its entirety. Be creative.

PRE-GENERATION STEPS:

- Identify all elements of the original song that relate to the definition of {MISOGYNY_TYPE}. These new elements must appear, in one form or another, in the newly generated version.
- Extract and argue the main themes, the genre as well as the mood of the next song.
- Generate a new song: Consider all of the above elements to generate a new song of acceptable quality that is sufficiently different from the original.

Once you have generated the new song, justify why it maintains the themes of the original song as well as the misogynistic elements. Please respond using the following outline:
<response>
<elements_m>Here you should write down all the misogynistic elements you have identified in the song along with their relationship to {MISOGYNY_TYPE}. Argue your answer by citing specific excerpts from the song that have allowed you to reach this conclusion. </elements_m>
<main_themes>Include here all the main themes surrounding the song with an argument for the presence of these"</main_themes>.
<mood>Mention here the main moods conveyed by this song</mood>.
<genre> Mention here the genre to which the song belongs</genre>.
<new_lyrics> The lyrics of the new song you have generated </new_lyrics>.
<justification>Justification of why the new version you have generated maintains the misogyny.
</justification>
</response>
Here are the lyrics to the song: {lyrics}