I2C-UHU at MiSonGyny 2025: Fine-Tuned LLM Approach to Detect Misogyny in Spanish Song Lyrics

Antonio Toro-Jaén, Victoria Pachón-Alvarez, Jacinto Mata-Vázquez and Angel Barroso-Romero

I2C Research Group, University of Huelva, Spain

Abstract

This work presents the approach developed by the I2C Research Group for the MiSonGyny 2025 shared task at IberLEF 2025, which focused on the detection of misogynistic content in Spanish-language song lyrics. The main contribution consists in the use of prompt-based fine-tuning applied to large language models (LLMs), combined with a parameter-efficient adaptation using LoRA. For Task 1 (binary classification), several LLMs were evaluated and integrated into an ensemble model based on majority voting, with tie-breaking performed using cumulative F1-scores. For Task 2 (multiclass classification), a single model was fine-tuned to classify lyrics into four categories of misogynistic content. The proposed system was evaluated on both tasks, achieving F1-scores of 83.59% and 56.13%, respectively, and obtaining the second-best results in both cases.

Keywords

Large language models (LLMs), LoRA, Ensemble, Prompt-based fine-tuning, Misogynistic content,

1. Introduction

Detecting hate speech in online and cultural content is a growing challenge in natural language processing (NLP), especially when such content is embedded in creative forms like music. Song lyrics, in particular, present complex linguistic constructions, figurative language, and subtle messages making it particularly difficult to accurately detect misogynistic messages. Music has a strong influence on public perceptions and social dynamics. When misogynistic language is present in lyrics, it can contribute to the reinforcement and normalization of harmful gender stereotypes. Identifying and addressing such content is therefore essential to promote more inclusive and respectful representations.

This paper describes our participation in the MiSonGyny 2025 [1] task at IberLEF 2025 [2], focused on the detection of misogynistic content in Spanish-language song lyrics. The task was divided into two subtasks: Task 1, which involved binary classification (misogynistic vs. not related), and Task 2, which required multiclass classification of the type of misogynistic expression (sexualization, violence, hate, or not related).

Recent advances in Large Language Models (LLMs) [3] and prompt-based training [4] have opened new possibilities for adapting models to highly contextual and nuanced tasks. This paper presents our participation in MiSonGyny 2025, where we applied LLM-based techniques to address both subtasks of the challenge.

2. Related Works

Recent studies have shown that song lyrics can reinforce gender-based stereotypes and sexist attitudes. As music is a widely consumed cultural product, the messages embedded in lyrics may influence social perceptions of gender roles. In recent years, researchers have applied natural language processing (NLP) techniques to detect sexism and bias in lyrics and to measure their evolution over time.

A notable example is the work by Casanovas-Buliart (2024) [5], who conducted a large-scale analysis of over 2,800 of the most-listened-to songs in Spain from 1960 to 2022. They used supervised machine

IberLEF 2025, September 2025, Zaragoza, Spain

antonio.toro@alu.uhu.es (A. Toro-Jaén); vpachon@dti.uhu.es (V. Pachón-Alvarez); mata@dti.uhu.es (J. Mata-Vázquez); angel.barroso880@alu.uhu.es (A. Barroso-Romero)

learning models trained on manually annotated paragraphs to classify songs as sexist or non-sexist. Their findings revealed that 51% of the songs contained sexist expressions, and that the presence of such content has increased significantly in recent years, especially in Spanish-language songs available on streaming platforms. They also identified dominant themes such as hypersexualization, objectification, and control over women.

Chen et al. (2025) [6] applied topic modeling (BERTopic) and gender bias metrics (SC-WEAT) to over 500,000 English lyrics. Their results showed that lyrics have increasingly associated women with terms related to appearance and sexuality, while men were more often linked to power and agency, highlighting a systematic gender imbalance in musical discourse.

Calderón-Suárez et al. (2023) [7] proposed a data augmentation strategy using misogynistic fragments from song lyrics to improve misogyny detection systems in social media. Their results suggest that music lyrics contain rich and diverse expressions of misogyny that can support the development of more robust detection models, particularly for Spanish-language data.

Beyond lyrics-focused work, related shared tasks have addressed sexism and misogyny detection in user-generated content. AMI @ EVALITA 2020 [8] targeted misogyny and aggressiveness in Italian tweets. Similarly, EXIST 2022 [9] offered multilingual benchmarks in English and Spanish for both binary and fine-grained sexism classification on social media, emphasizing the complexity of capturing implicit forms of bias.

Building on these trends, the HOMO-MEX 2024 [10] included a subtask specifically focused on detecting LGBT+phobic content in Spanish song lyrics. The dataset combined over 1,200 annotated lyrics using both automatic methods (via the Genius API and LyricScraper) and manual selection informed by members of the LGBT+ community. The task highlights how challenging it is to detect hate speech when it appears in figurative or poetic language, since these forms often hide harmful content behind metaphors or artistic expressions. Successfully identifying them requires models capable of deep contextual understanding and subtle interpretation. In contrast to prior studies that focus on general sexism or use lyrics as a resource for other tasks, the present work addresses the direct detection of misogyny in Spanish song lyrics. We apply a fine-tuned large language model to perform both binary and multiclass classification, aiming to bridge the gap between NLP methods and the cultural analysis of lyrical content.

3. Task and Dataset Description

The MiSonGyny 2025 challenge includes two independent classification tasks focused on detecting and categorizing misogynistic content in Spanish-language song lyrics. Each task is framed as a supervised learning [11] problem using manually labeled datasets.

3.1. Task 1: Binary Classification

The objective of Task 1 is to classify each song lyric as either misogynistic or non-misogynistic. The definitions for the two labels are as follows:

- **Misogynistic (M)**: Lyrics that include explicit or implicit content that denigrates, disrespects, or expresses hatred, sexual objectification, or control towards women.
- **Non-misogynistic (NM)**: Lyrics that do not contain any such content, regardless of whether or not they refer to gender or women.

A dataset of 2,105 entries was provided for this task, all in Spanish, with some containing short English phrases. The dataset was manually split in two stages:

1. An 80/20 split was performed to separate the full dataset into training and test sets.
2. The training portion (80%) was the further split 80/20 into a new training set and a validation set.

The final distribution of classes across the splits is shown in Table 1.

Table 1
Classes distribution for Task 1

Class	Train Dataset	Valid Dataset	Test Dataset
NM	935	234	293
M	411	103	128
Total	**1346**	**337**	**421**

3.2. Task 2: Multiclass Classification

The second task is a four-class classification problem. Each lyric was labeled according to the specific type of misogynistic expression:

- **Sexualization (S)**: Expressions that imply or mention sexual behavior, use sexual terms, or contain suggestive remarks.
- **Violence (V)**:Expressions that involve threats, acts of physical harm, or verbal abuse.
- **Hate (H)**: Expressions that use offensive or biased language, or convey disdain or antagonism toward individuals or groups.
- **Not Related (NR)**: Expressions that don't fit into the other categories and contain no sexual, violent, or hateful content.

This dataset contains 1,168 lyrics, and was split using the same two-stage stratified strategy:

1. 80% for training and 20% for testing.
2. Then 80%/20% split on the training data for the validation data.

The final distribution of classes across the splits is shown in Table 2.

Table 2
Classes distribution for Task 2

Class	Train Dataset	Valid Dataset	Test Dataset
NR	337	84	105
S	278	70	87
V	82	21	26
H	50	12	16
Total	**747**	**187**	**234**

4. Methodology

This work explores the use of Large Language Models (LLMs) [12] to detect and classify misogynistic content in Spanish-language song lyrics. The methodology was designed to address two distinct tasks: a binary classification (Task 1) and a multiclass classification (Task 2), both framed as supervised text classification problems.

To solve these tasks, several open-access LLMs were fine-tuned using a parameter-efficient approach based on Low-Rank Adaptation (LoRA). [13] The training process included custom output constraints and data balancing strategies to adapt the models to the specific nature of lyric-based misogyny detection.

All models were trained using a NVIDIA GeForce RTX 4070 GPU with 12 GB of VRAM. While this setup allowed for efficient experimentation, the available memory, training time, and computational budget imposed significant constraints. Without the use of parameter-efficient fine-tuning techniques such as LoRA and 4-bit model quantization, it would not have been feasible to train and evaluate 8B-scale language models. These strategies were therefore essential to reduce both memory usage and training time, enabling the implementation of multiple model variants within the available resources.

4.1. Data preprocessing

Before training, several preprocessing steps were applied to clean and normalize the data. All lyrics were first lowercased, and irrelevant content, such as punctuation marks, artist names, song titles, and non-lyrical text, was removed. Table 3 shows an example before and after preprocessing.

Table 3
Example of original and preprocessed lyrics used in the training pipeline.

Original	Preprocessed
[Intro-Scratch] Flor de vida, flor de vida... Los paisajes cambian Un chico de carácter agradable [?] de suicidio, puede grabarlo en vídeo ...	flor de vida flor de vida los paisajes cambian un chico de carácter agradable de suicidio puede grabarlo en vídeo

For Task 1, the original class labels were 'M' (misogynistic) and 'NM' (non-misogynistic). For Task 2, the labels were 'NR' (not related), 'S' (sexualization), 'H' (hate), and 'V' (violence). All labels were subsequently mapped to integer values.

To address class imbalance, random oversampling was applied to the training set in both tasks. In Task 1, the original distribution consisted of 935 non-misogynistic and 411 misogynistic samples in the training set. After oversampling, the minority class was increased to 600 samples, reducing the imbalance while avoiding excessive duplication. In Task 2, the initial distribution was 337 for 'NR', 278 for 'S', 82 for 'V', and 50 for 'H'. After oversampling, the 'V' and 'H' classes were increased to 150 and 100 samples respectively, achieving a more balanced representation without fully equalizing all classes.

4.2. Model Selection and Training Setup

To solve both tasks, Large Language Models (LLMs) were used with multilingual capabilities and support for long-context input. All models were used in their bnb 4-bit quantized versions for efficient fine-tuning using Unsloth [14], which enables fast loading and compatibility with LoRA-based adaptation. The selected models were:

- **DeepSeek-R1-Distill-Llama-8B-unsloth-bnb-4bit:** A distilled version of DeepSeek-R1, fine-tuned on LLaMA 3.1 outputs. It offers strong reasoning performance with reduced computational demands. [15]
- **Meta-Llama-3.1-8B-bnb-4bit:** Meta's multilingual LLM optimized for alignment and dialogue, with robust performance in Spanish and other supported languages. [16]
- **Qwen3-8B-Base-unsloth-bnb-4bit:** Alibaba's third-generation model, designed for strong instruction-following and multilingual capabilities, with high efficiency at 8B scale. [17]

All three models were evaluated in Task 1, while only Qwen3 was used for Task 2 due to its superior performance in the binary classification task.

Training Configuration

Different training configurations were applied depending on the model. For DeepSeek-R1, fine-tuning was carried out for 5 epochs with an effective batch size of 8. The training set contained 1,526 examples, resulting in approximately 190 steps per epoch and a total of 950 training steps. The learning rate was set to 2e-5, and the optimizer used was 8-bit AdamW. The training process included early stopping with a patience of 3 and checkpoint saving every 25 steps. LoRA was configured with rank = 16 and alpha = 16.

For Qwen3 and LLaMA 3.1, training was performed for 10 epochs in both tasks where they were used, with an effective batch size of 64. The learning rate remained at 2e-5, using the same LoRA

configuration as above. The optimizer used was also 8-bit AdamW. A linear learning rate schedule was applied, and early stopping based on validation loss was enabled with a patience of 3. Checkpoints were saved every 25 steps, and the best model was restored automatically based on evaluation loss. In Task 2, only Qwen3 was used, and it followed the same configuration as in Task 1.

Prompting Strategy

Different prompt templates were designed for each model to optimize classification performance. For **DeepSeek-R1**, an instruction-based prompt was used that explicitly listed the criteria for identifying misogynistic content and encouraged the model to reason step by step using a chain of thought format. The exact prompt is shown below:

> ### Instruccion
>
> Analiza la siguiente letra de canción y determina si contiene contenido misógino. Evalúa si incluye lenguaje, actitudes o mensajes que:
> - Degraden o deshumanicen a las mujeres.
> - Menosprecien a las mujeres de manera explícita o implícita.
> - Refuercen estereotipos negativos o dañinos sobre las mujeres.
> - Promuevan violencia física, emocional o sexual contra las mujeres.
>
> Piensa cuidadosamente tu respuesta y crea paso a paso una chain of thoughts para dar una respuesta logica.
> Responde únicamente con "1" si la letra es misógina o con "0" si la letra no es misógina. No proporciones ninguna explicación ni texto adicional.
>
> ### Letra:
> `{lyrics}`
>
> ### Respuesta:
> `<think>{reasoning}</think> {label}`

For **Qwen3** and **LLaMA 3.1**, classification was prompted using a more direct and concise template:

> Analiza la siguiente letra de canción: `{lyrics}`
>
> Clasifica esta canción en una de las siguientes categorías:
> Clase 1: No Misógina
> Clase 2: Misógina
> **SOLUCION**
> La cancion es: Clase `{label}`

Additionally, for Task 2, where lyrics had to be classified into four categories, we used the following instruction:

> Analiza la siguiente letra de canción: `{lyrics}`
>
> Clasifica esta canción en una de las siguientes categorías:
> Clase 1: No Relacionada
> Clase 2: Sexualizacion
> Clase 3: Violencia
> Clase 4: Odio
> **SOLUCION**
> La cancion es: Clase `{label}`

In addition, a more detailed variant of the prompt was tested, including explicit descriptions of each class and examples of misogynistic content. However, this version led to slightly worse performance, likely due to increased token length and greater generation variability. To ensure consistency and prevent unexpected outputs, the lm_head layer of Qwen3 and LLaMA was modified to restrict predictions

to only the token IDs corresponding to the allowed class labels (1 and 2 for Task 1, and 1 to 4 for Task 2). Furthermore, the training setup was adapted so that only the final token in the sequence contributed to the loss function, ensuring that the model focused solely on predicting the class label rather than attempting to reconstruct the entire input. This adjustment improved both training efficiency and classification stability.

Tokenization

All models were used with their corresponding tokenizers. Given the nature of song lyrics, which often include repeated phrases, annotations, or lengthy verses, special attention was paid to input length. A maximum sequence length of 4,096 tokens was defined. Lyrics exceeding the limit were truncated from the end of the sequence, preserving the start of the prompt and lyric. This ensured uniform input size across the dataset.

Reproducibility

To ensure reproducibility, all experiments were conducted using the same random seed (3407) and consistent dataset splits. The code used for model training is available at:

https://github.com/atoroj/Misogyny-Lyrics-Detection-LLMs

This repository includes the notebooks, dataset, and everything necessary to reproduce the results presented in this work.

4.3. Models Performance

This section presents the final evaluation results for both tasks as part of the official MiSonGyny 2025 competition. After training and selecting the best-performing models using test data extracted from the provided training set, the predictions for each task were submitted. The results were evaluated using official competition metrics, macro-averaged F1 score, precision and recall, as shown in Table 4.

Table 4
Evaluation metrics (F1, Precision, Recall).

Model	F1-Score	Precision	Recall
Task 1			
DeepSeek-R1 8B	0.72022	0.77090	0.70224
Llama3.1 8B	0.81209	0.81630	0.80833
Qwen3 8B	0.83592	0.83774	0.83417
Task 2			
Qwen3 8B	0.56131	0.57093	0.55614

4.3.1. Ensemble

To improve performance in Task 1, an ensemble strategy was explored by combining the predictions of four independently fine-tuned models. The ensemble output was determined by majority voting over the predictions of the four models. If three or more agreed on the same label, that prediction was accepted. [18]

In cases of a 2 vs 2 tie, the resolution strategy was based on a weighted F1-score voting. Specifically, the sum of the individual F1-scores of the models voting for each class was computed, and the class with the higher cumulative F1 was selected.

Table 5 shows the F1-scores obtained by each individual model, as well as the final F1-score achieved by the ensemble.

Table 5
Ensemble of models.

Model	F1-Score
Qwen3-8B (Prompt 1)	0.8359
Qwen3-8B (Prompt 2)	0.8137
Llama3.1-8B (Prompt 1)	0.8121
Qwen3-14B (Prompt 1)	0.8224
Ensemble	0.83201

Although the ensemble did not outperform the best individual model, it still achieved a highly competitive F1-score and improved the stability of predictions.

5. Results

The final evaluation results are presented in Table 6 and Table 7 . The table displays the usernames, F1-scores, and ranking positions for the top five participants in each task. All metrics were calculated over a hidden test set using official evaluation scripts.

In Task 1 (Binary Classification), our system obtained an F1-score of 0.8359, ranking 2nd place overall. Our approach demonstrated solid and consistent performance across all instances.

Table 6
Task 1: Binary Classification – Top 5 Rankings

Rank	Username	F1-score
1	jstoledo	0.88114
2	**I2C-Vega (Ours)**	**0.83592**
3	carlosdf	0.82795
4	mapachepunk	0.80514
5	eliasurios30	0.80394

In Task 2 (Multiclass Classification), which involved detecting the specific type of misogynistic content, our model achieved an F1-score of 0.5613, also placing 2nd among all submissions.

Table 7
Task 2: Multiclass Classification – Top 5 Rankings

Rank	Username	F1-score
1	jstoledo	0.58954
2	**I2C-Vega (Ours)**	**0.56131**
3	eliasurios30	0.54299
4	luisramos07	0.49462
5	danielacl	0.49287

The strong performance across both tasks highlights the effectiveness of the proposed methodology. Our model ranked above other systems, demonstrating the relevance of model selection, careful prompting, and parameter-efficient fine-tuning in this type of linguistic and socially sensitive classification.

6. Error Analysis

The confusion matrix for Task 1 is shown in Figure 1. The model demonstrates a strong ability to detect non-misogynistic lyrics (Class 0), with 267 true negatives and only 26 false positives. However,

31 misogynistic lyrics (Class 1) were misclassified as non-misogynistic, revealing a vulnerability in detecting subtle or implicit forms of misogyny. These false negatives often involved figurative language, metaphorical expressions, or indirect verbal aggression that the model failed to interpret as harmful.

On the other hand, some false positives appear to stem from lyrics expressing passion, sexuality, or emotional intensity, particularly from a female perspective. This suggests that the model may occasionally confuse female empowerment or sensual expression with misogynistic stereotypes.

Table 8 illustrates selected examples of misclassified lyrics from Task 1.

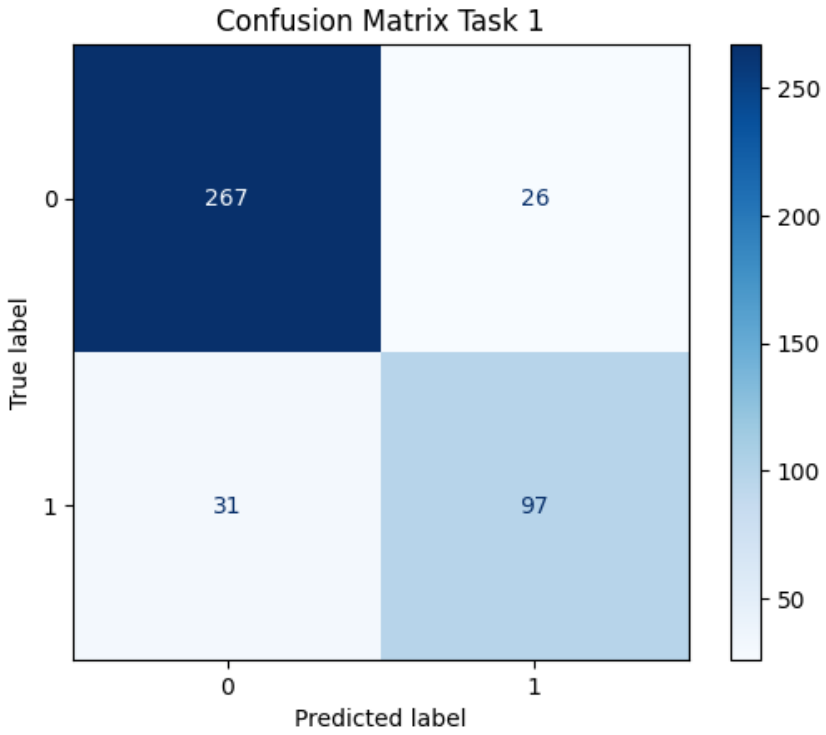

Figure 1: Confusion matrix for Task 1, 0 for No Misogynistic and 1 for Misogynistic.

Table 8
Examples of misclassified lyrics.

Lyrics	Predicted Label	True Label
"Quiere que lo hagamos en la discoteca... la G te come, pero no te besa..."	M	NM
"Si tú quieres un atajo y lo quieres por abajo, yo te llevo bien callao..."	M	NM
"Haber sido siempre tu abrigo yo debí ser más dura contigo y dejar que murieras de frío..."	NM	M
"Si me provocas te voy a besar los ojos te voy a tomar del pelo te voy a hacer llorar de un beso..."	NM	M

Figure 2 shows the confusion matrix for Task 2. It shows that the model performs reasonably well in distinguishing Not Related lyrics (Class 0) and Sexualization (Class 1), with 80 and 62 correct predictions, respectively. But significantly lower performance in identifying Violence (Class 2) and Hate (Class 3). Only 6 and 4 instances were correctly classified, respectively, while many others were predicted as Class 0 or Class 1. This may be attributed to the data imbalance affecting the Violence and Hate classes, as well as the semantic overlap among hate speech, violent expressions, and sexualized content.

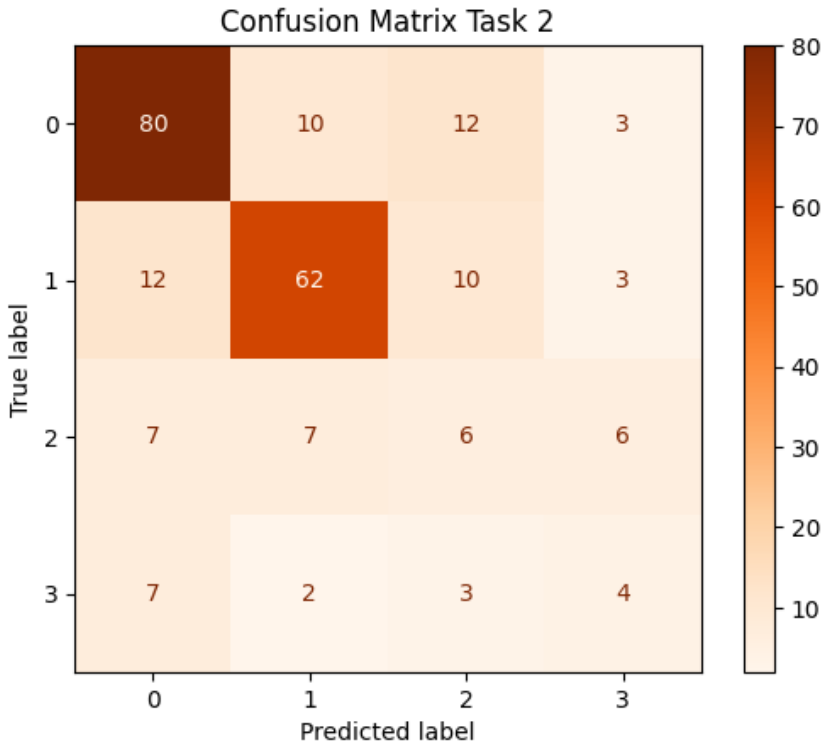

Figure 2: Confusion matrix for Task 2, 0 for Not Related, 1 for Sexualization, 2 for Violence and 3 for Hate.

7. Conclusion

In this paper, we presented our system for the IberLEF 2025-MiSonGyny task, which addressed the detection of misogynistic content in Spanish-language song lyrics. We participated in both Task 1 (binary classification) and Task 2 (multiclass classification), achieving robust and competitive results in both. Our approach was based on fine-tuning multilingual Large Language Models (LLMs), with prompt-based fine-tuning.

These findings highlight the potential of LLMs trained with prompt-based fine-tuning to learn task behavior efficiently, especially when provided with well-structured input formats. In future works, we plan to explore larger or instruction-tuned models to enhance generalization, apply data augmentation for underrepresented classes, and investigate the use of ensemble methods to further improve performance.

In conclusion, our study shows that using carefully designed prompts to fine-tune large language models is an effective way to address complex classification tasks, especially when working with creative and culturally rich texts such as song lyrics. The results highlight how a careful combination of good data preparation, efficient training, and focused evaluation can lead to strong performance on real-world language challenges.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o in order to: Text Translation, Grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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).

[2] 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.

[3] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, 2024. URL: https://arxiv.org/abs/2307.06435. arXiv:2307.06435.

[4] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. URL: https://arxiv.org/abs/2107.13586. arXiv:2107.13586.

[5] 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.

[6] D. Chen, A. Satish, R. Khanbayov, C. M. Schuster, G. Groh, Tuning into bias: A computational study of gender bias in song lyrics, 2025. URL: https://arxiv.org/abs/2409.15949. arXiv:2409.15949.

[7] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, M. A. Márquez-Vera, Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases, IEEE Access 11 (2023) 13179–13190. doi:10.1109/ACCESS.2023.3242965.

[8] P. R. Elisabetta Fersini, Debora Nozza, Ami @ evalita2020: Automatic misogyny identification, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[9] F. R.-S. y Jorge Carrillo-de-Albornoz y Laura Plaza y Adrián Mendieta-Aragón y Guillermo Marco-Remón y Maryna Makeienko y María Plaza y Julio Gonzalo y Damiano Spina y Paolo Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[10] 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.

[11] S. B. Kotsiantis, Supervised machine learning: A review of classification techniques, Informatica 31 (2007) 249–268.

[12] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[14] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.

[15] D. AI, Deepseek llms, https://huggingface.co/DeepSeek-AI, 2025.

[16] M. Llama, Llama ai, https://huggingface.co/meta-llama, 2025.

[17] A. Qwen, Qwen llms, https://huggingface.co/Qwen, 2025.

[18] G. Abramowitz, Model independence in multi-model ensemble prediction, Australian Meteorological and Oceanographic Journal 59 (2010) 3–6. doi:10.22499/2.5901.002.