# ELiRF-UPV at MisonGyny-IberLEF 2026: Misogyny in Song Lyrics

Vicent Ahuir[2,†], Sergio Marín Muñoz[1,†] and María José Castro-Bleda[2,3,*,†]

[1]*Department of Computer Systems and Computation, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

[2]*VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

[3]*ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

Abstract

The detection of misogyny in song lyrics introduces unique linguistic and cultural challenges, including implicit bias, genre-specific tropes, and contextual ambiguity. In this working note, we report our participation in the MiSonGyny shared task across all three official tracks. We describe our chunking-based preprocessing pipeline, domain-adapted model selection, and task-specific training strategies. Our system achieves an overall ranking of 9th-10th place, with Task 1, Task 2, and Task 3 Macro-F1 scores of 0.8334, 0.6562, and 0.7724, respectively. We discuss key technical decisions, failure modes, and cross-task generalization patterns, and outline directions for future work in culturally grounded bias detection.

**Keywords**

Misogyny Detection, Song Lyrics Analysis, Natural Language Processing, Transformers-based Models, Cultural Text Mining

## 1. Introduction

Misogyny, understood as any form of discrimination, prejudice, or aversion towards women based on their gender, is a deeply rooted global issue spanning generations. The consequences women face due to misogynistic attitudes include gender-based violence, economic inequalities, social exclusion, mental health issues, and unequal access to education and employment opportunities. This problem manifests in personal and social interactions, as well as in popular culture, particularly in songs and television shows. Music, as a massive form of cultural expression, often perpetuates negative stereotypes about women through lyrics that objectify, sexualize, or demean them. This directly influences how society normalizes such behaviors and reinforces misogynistic attitudes.

Misogyny in cultural expressions poses an additional challenge for automated systems. The language used in songs and television content is often ambiguous or disguised as humor, satire, or artistic style, making automatic detection difficult. This highlights the need for NLP models

*IberLEF 2026, September 2026, León, Spain*

*Corresponding author.

†These authors contributed equally.

vahuir@dsic.upv.es (V. Ahuir); smarmuo@etsinf.upv.es (S. M. Muñoz); mcastro@dsic.upv.es (M. J. Castro-Bleda)

0000-0001-5636-651X (V. Ahuir); 0000-0003-1001-8258 (M. J. Castro-Bleda)

to address inherent biases in training data and incorporate a deeper understanding of cultural context to adequately identify these forms of misogyny. It also reveals a structural bias in deep learning systems, which often fail to distinguish between legitimate complaints about discrimination and openly misogynistic language. Ironically, this technological limitation can end up silencing women's voices and perpetuating the very inequalities NLP models should mitigate.

The MiSonGyny task addresses a poorly studied area of Spanish-language song lyrics, where the apology or promotion of hate speech against women is frequently present. This benchmark enables participants to explore underutilized linguistic resources and develop more precise methods for detecting misogynistic content in culturally rich, structurally complex textual domains.

## 2. Task Overview

Recent shared tasks have increasingly focused on the automated analysis of toxic discourse in Spanish music, such as the HOMO-MEX task at IberLEF 2024 [1], which established a benchmark for LGBT+ hate speech detection in song lyrics. Expanding on this research line, the MiSonGyny 2026 shared task [2] at IberLEF 2026 [3] follows the 2025 edition [4] to focus specifically on the structural and semantic identification of misogyny. We participated in all three MiSonGyny tracks:

- **Task 1:** Binary classification of misogynistic (M) vs. non-misogynistic (NM) lyrics.
- **Task 2:** Multilabel discrimination of explicit manifestations: sexualization (S), violence (V), and hate (H).
- **Task 3:** Binary classification of stereotypes (Yes/No).

A critical annotation characteristic to note is that Tasks 2 and 3 were only labeled for lyrics already classified as misogynistic (M) in Task 1. This conditional labeling scope impacts class distribution and evaluation context. All tracks are evaluated using Macro-F1. Our team submission ID: `sergiomarinnx`.

### 2.1. Dataset Summary

Firstly, an analysis of song distribution across the training dataset was conducted to understand the data used to train the models. A balanced proportion of misogynistic and non-misogynistic songs is observed, accounting for 56.2% and 43.8%, respectively, in Task 1. Regarding the song location, there are two different categories in the dataset, from Spain and from Latin America, labeled `es_ES` and `es_LATAM`, with 1181 and 1122 instances, respectively. Approximately 55% of the misogynistic songs originate from Latin America, while 52% of the non-misogynistic songs are from Spain, indicating a slight correlation between the origin of the song and the misogynistic factor.

Another important point is the appearance of anglicism in these songs, which is more commonly found in the reggaeton, rap, and trap musical genres. A vocabulary study revealed that words such as *yeah*, *baby*, *party*, *bitch*, and *love* are more strongly associated with misogynistic songs.

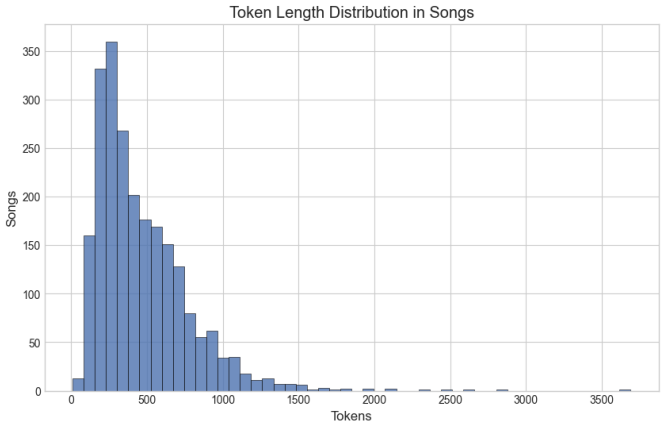


**Figure 1:** Token length distribution for all songs.

### 2.2. Dataset Analysis

The primary challenge across all tasks was the length of song lyrics, which frequently exceed the 512-token context window of standard BERT-style architectures. To characterize this domain property, we analyzed the token length distributions across the full dataset and its task-specific subsets, as shown in Figures 1–4.

Figure 1 presents the overall distribution, revealing a pronounced right skew with a significant portion of songs extending beyond 500 tokens. This confirms that naive truncation would discard substantial narrative and poetic context. Task 1 analysis (Fig. 2) further highlights this effect: misogynistic (M) tracks exhibit a longer tail compared to non-misogynistic (NM) tracks, with many extending past 800 tokens. This class-specific distribution directly motivated our overlapping chunking strategy, which preserves contextual integrity without arbitrary truncation.

For the conditional tasks, Figures 3 and 4 show the distributions for Task 2 (Sexualization, Violence, Hate) and Task 3 (Stereotypical, Non-Stereotypical), respectively. Notably, Task 2 samples concentrate in the 500–1000 token range, suggesting that while these lyrics are shorter on average, they carry a dense semantic load. The restricted annotation scope (Task 2 and Task 3 labels only apply to pre-flagged M tracks) also influences the observed distributions, as longer, more narrative-driven songs are more likely to be annotated with fine-grained bias categories. Together, these distributions validate our preprocessing pipeline: metadata-aware chunking with 250-token windows and 50-token overlaps ensures that both short, dense lyrical segments and longer narrative passages are processed with minimal context loss.

## 3. Methodology

### 3.1. Preprocessing and Contextualization

To preserve lyrical structure, narrative flow, and cultural context, we implemented the following pipeline:

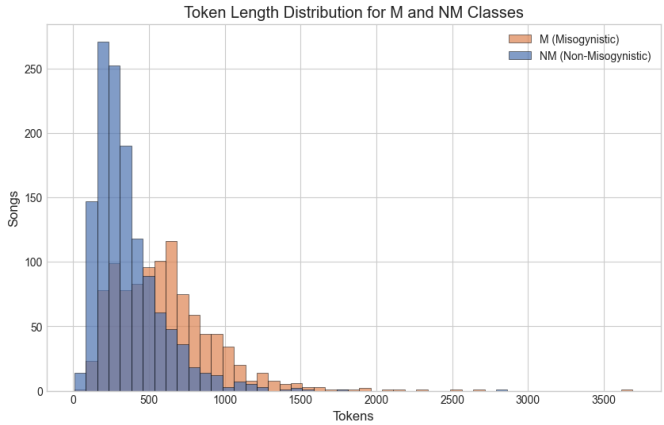


**Figure 2:** Token length distribution for Task 1: Misogynistic (M) vs. Non-Misogynistic (NM) classes.

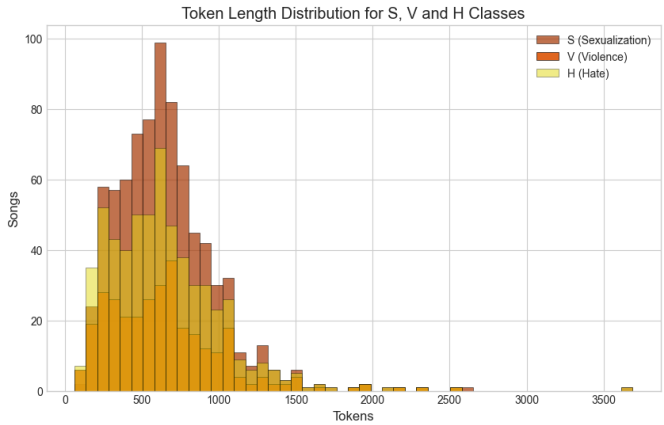


**Figure 3:** Token length distribution for Task 2: Sexualization (S), Violence (V), and Hate (H).

- **Metadata Injection:** We prefixed each lyric with the artist's regional localization token ('es_LATAM' or 'es_ES') followed by the song title, separated by '[SEP]'. This provided explicit geographic and cultural grounding crucial for disambiguating regional slang, artistic tropes, and contextual references.
- **Chunking Strategy:** Lyrics were split into overlapping chunks of 250 tokens with a 50-token overlap. This preserved sentence boundaries, avoided truncating poetic phrasing, and maintained local context. The tokenizer's maximum length was fixed at 300 tokens to accommodate the overlap while staying within model limits.
- **Pooling Mechanism:** Chunk-level predictions were aggregated using *Mean-Pooling* across token embeddings before the classification head. This consistently outperformed Max-Pooling ablations.

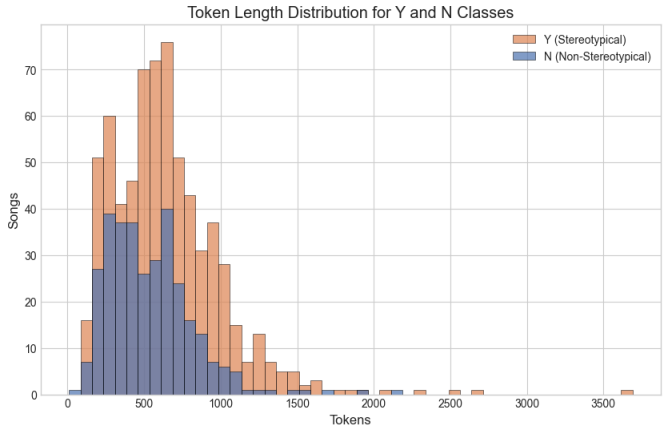


**Figure 4:** Token length distribution for Task 3: Stereotypical (Y) vs. Non-Stereotypical (N).

### 3.2. Model Selection and Training Configuration

We evaluated BETO [5] and XLM-RoBERTa [6], with BETO yielding superior results across all tracks, likely due to its optimization for Spanish morphosyntax and lexical distributions. All models were trained on a consumer laptop equipped with an NVIDIA RTX 4060 GPU. To manage memory constraints, we used a per-device batch size of 2 with `gradient_accumulation_steps=4`, simulating an effective batch size of 8.

We explored multiple training variants (hyperparameters, class weighting, pooling strategies). The highest-performing configuration was *Version 3* across all tasks. Task-specific configurations are detailed below:

**Task 1 (Binary M/NM):** BETO, `test_size=0.2`, `num_train_epochs=4`, `learning_rate=2e-5`, `warmup_ratio=0.1`, `weight_decay=0.01`, batch=2, grad_acc=4. Classification head: Mean-Pooling + Softmax. F1: 0.8334.

**Task 2 (Multilabel S/V/H):** BETO, `test_size=0.2`, `num_train_epochs=3`, `learning_rate=2e-5`, `warmup_ratio=0`, `weight_decay=0.01`, batch=2, grad_acc=4. Classification head: Mean-Pooling + Sigmoid. F1: 0.6562.

**Task 3 (Binary Stereotypes):** BETO, `test_size=0.2`, `num_train_epochs=4`, `learning_rate=2e-5`, `warmup_ratio=0.1`, `weight_decay=0.01`, batch=2, grad_acc=4. Classification head: Mean-Pooling + Softmax. F1: 0.7724.

An ablation in *Version 4* introduced class weights and switched to Max-Pooling, which degraded performance to F1=0.7609, confirming that Mean-Pooling better captures the distributed semantic signals in lyrical chunks.

## 4. Results and Analysis

Table 1 summarizes our performance across the three tasks, comparing our primary chunk-aware BETO pipeline (Version 3) against an XLM-RoBERTa baseline (Version 5). We achieved an overall ranking of 9th-10th place among 11 participating teams.

**Table 1**
Performance comparison across models and tracks

| Task | Model | Description | Macro-F1 |
|---|---|---|---|
| 1 | BETO | Binary misogyny detection | 0.8334 |
| 1 | XLM-RoBERTa | Binary misogyny detection | 0.8076 |
| 2 | BETO | Multilabel S/V/H classification | 0.6562 |
| 2 | XLM-RoBERTa | Multilabel S/V/H classification | 0.4905 |
| 3 | BETO | Binary stereotype detection | 0.7724 |
| 3 | XLM-RoBERTa | Binary stereotype detection | 0.7647 |

**Key Observations:**

- **Task 1:** The strong F1 score ($0.8334$) reflects the model's capacity to leverage chunk-level context and metadata tokens to identify direct lexical and tonal markers of misogyny.
- **Task 2:** The performance drop ($0.6562$), which heavily influenced our 9th-10th overall placement, highlights the difficulty of fine-grained multilabel discrimination. A qualitative error analysis reveales that this lower performance is primarily driven by label overlap and semantic blurring. Implicit sexualization, contextual violence, and hate speech often overlap in lyrical metaphors, and the conditional annotation scope (only misogynistic tracks) limited positive sample diversity.
- **Task 3:** The $0.7724$ score indicates robust stereotype detection, though performance is sensitive to genre-specific tropes and evolving cultural norms around gender representation.
- **Pooling vs. Naive Truncation:** To validate the necessity of our sliding window strategy, an ablation baseline using naive truncation (statically cutting text at $250$-tokens) was evaluated. This baseline achieved Macro-F1 scores of $0.8331$, $0.5724$, and $0.7532$ for Tasks 1, 2 and 3, respectively. The consistent performance lift achieved by our chunking pipeline (Mean-Pooling across $250$-token chunks) across all tracks confirms that preserving full sub-document narratives prevents information loss.

## 5. Discussion and Lessons Learned

1. **Lyrical Context vs. Static Truncation:** Forcing lyrics to $512$ tokens truncates essential song structures (e.g., late-track choruses where toxicity often peaks). Our sliding window strategy proved that preserving full lyrical arcs is strictly necessary for domain-adapted misogyny detection, as evidenced by our baseline comparisons.
2. **Hardware Constraints and LLM Absence:** Our hardware-aware training pipeline (restricted to an RTX 4060 GPU) limited our architectural exploration to encoder-only models (BETO, XLM-R) rather than modern generative Large Language Models (LLMs). We hypothesize that our lower relative ranking (9th-10th place) is partially due to the inability of smaller encoders to perform the deep implicit reasoning required for subtle stereotype detection, a task where larger models excel.

3. **Multilabel Blurring Effect:** Task 2's lower score (0.6562) exposes a fundamental domain difficulty: explicit manifestations of misogyny in urban music are rarely mutually exclusive. For instance, aggressive sexualization in reggaeton frequently overlaps lexically with physical violence. Standard BCE loss functions struggle to cleanly separate these heavily entangled cultural tropes.
4. **Cultural and Linguistic Grounding:** Injecting regional tokens ('es_LATAM'/'es_ES') and titles significantly improved disambiguation, confirming that metadata-aware preprocessing is crucial for culturally situated NLP tasks.
5. **Evaluation Limitations:** We acknowledge the absence of statistical significance testing across our ablations. While Mean-Pooling consistently improved F1 scores empirically over Max-Pooling during our validation phases, future iterations must incorporate rigorous significance bounds to confirm these architectural decisions.

## 6. Conclusion and Future Work

Our participation in the MiSonGyny shared task demonstrates that chunk-aware preprocessing, metadata grounding, and task-specific pooling strategies yield competitive performance for lyrical misogyny detection. Achieving 9th place across all three tracks reflects a solid baseline but also reveals clear gaps in implicit bias reasoning, multilabel discrimination, and generalization under conditional annotation schemes. Future directions include:

- Adaptive chunking with semantic boundary detection (e.g., stanza/chorus markers)
- Hierarchical or prompt-based multilabel classification for Task 2
- Full-context annotation pipelines to reduce selection bias
- Integration of audio-prosodic or genre metadata for multimodal grounding
- Expanded evaluation on underrepresented Spanish dialects and independent music corpora

We thank the MiSonGyny organizers for building a thoughtfully structured benchmark and for fostering critical NLP research at the intersection of ethics, culture, and digital media.

## Declaration on Generative AI

During the preparation of this work, the author(s) used LLMs in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

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

[2] T. Alcántara, E. Urios Alacreus, C. Macías, M. Aloy Mayo, O. Garcia-Vazquez, A. Espinosa-Juarez, M. Soto, P. Rosso, J.-E. Valdez-Rodríguez, H. Calvo, Overview of MiSonGyny at IberLEF 2026: Misogyny Speech Detection in Mexican and Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural To be announced (2026).
[3] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.
[4] 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).
[5] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, BETO: Spanish Pre-trained Language Model, https://arxiv.org/abs/2005.11045, 2020. arXiv:2005.11045.
[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

## Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and ERDF "ERDF A way of making Europe" (project PID2024-155948OB-C55).