aerojash at MiSonGyny 2026: A Hierarchical Approach to Misogyny Detection in Spanish Song Lyrics

Alix Erica Rojas-Hernández[1,*], Luis Fernando Niño-Vásquez[1] and Fabio Augusto González-Osorio[1]

[1]*Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia*

Abstract

Misogyny in song lyrics is often expressed through subtle stereotypes, possessive language, sexualization, and symbolic violence rather than explicit insults, making automatic detection particularly challenging. MiSonGyny 2026, the IberLEF shared task on misogyny detection in Spanish song lyrics, addresses this problem through three hierarchical subtasks: binary misogyny detection, misogyny-type classification, and stereotype identification. We present the aerojash submission, a gated multi-task system that combines a multilingual encoder with task-specific prediction heads, a first/last token-selection strategy that concatenates tokens from the beginning and end of over-length lyrics, and a deterministic inference gate in which the binary misogyny decision constrains downstream predictions of type and stereotype. Our approach ranked 3rd of 11 teams in misogyny detection (macro-F1 = 0.858) and 3rd of 11 in stereotype identification (0.807), while achieving 0.668 in misogyny-type classification. Beyond the official results, we analyze the behavior of the classifier as a working notes case study in hierarchical multi-task classification. The analysis documents substantial differences associated with thresholding, hierarchical inference, and validation-time decisions, although their individual effects cannot be isolated. These observations underscore the need to report inference and validation decisions explicitly in shared tasks involving long, stylistically diverse lyrics.

Keywords

Misogyny detection, Spanish song lyrics, Hierarchical multi-task learning, XLM-RoBERTa, IberLEF, MiSonGyny 2026

1. Introduction

Misogyny in song lyrics is rarely expressed as explicit hate. It surfaces through sexualization, possessive address, symbolic violence, gender stereotypes, repeated hooks, slang, and genre-specific conventions that are difficult to interpret, even for human annotators. This makes lyrics a demanding testbed for hate-speech detection: documents are long, stylistically dense, culturally situated, and frequently ambiguous regarding whether misogynistic expressions reflect storytelling, artistic performance, or genuine endorsement of misogynistic ideas.

The MiSonGyny 2026 shared task [1], held at IberLEF 2026 [2], operationalizes this challenge through three song-level subtasks: binary misogyny detection (T1), where songs are classified as Misogynistic (M) or Non-Misogynistic (NM); multilabel categorization of misogynistic content (T2) into Sexualization (S), Violence (V), and Hate (H); and stereotype identification (T3), where songs are labeled as containing stereotypes (Y) or not (N). The task is hierarchical by design: songs predicted as NM receive no T2 labels and are assigned N in T3 by construction. Systems must therefore not only identify each phenomenon but also maintain consistency across interdependent predictions.

Participating under the team alias aerojash, we submitted a gated multi-task system based on XLM-RoBERTa-large [8]. The model encodes each lyric with a first/last token-selection strategy for long documents, shares the encoder across three prediction heads trained jointly under Focal Loss [9], and applies a deterministic T1 gate at inference time. The design is deliberately simple and well aligned

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

aerojash@unal.edu.co (A. E. Rojas-Hernández)

0000-0002-0371-3925 (A. E. Rojas-Hernández); 0000-0003-4703-0007 (L. F. Niño-Vásquez); 0000-0001-9009-7288 (F. A. González-Osorio)

with the task structure: T1 decides whether the downstream misogyny-specific heads may fire, while T2 and T3 specialize in content type and stereotype prediction. The final submission ranked 3rd/11 on T1 (macro-$F_1 = 0.8581$) and 3rd/11 on T3 (0.8071), with a lower score on T2 (0.6675); the subsequent development-set error analysis identifies violence as a recurring source of difficulty. Beyond presenting our approach, we examine several behaviors that emerged during development and evaluation. Three observations motivated a closer analysis: an unusually low operating threshold for stereotype detection, an apparent gap between development and test performance for type classification, and the failure of a development-selected ensemble to retain its advantage on the official test set.

The remainder of the paper is organized as follows. Section 2 introduces the task and dataset. Section 3 describes the system architecture and training strategy. Section 4 reports the official results and analyzes the main development-time observations. Section 5 presents the error analysis, and Sections 6 and 7 discuss limitations and conclusions.

2. Related Work and Task Setup

2.1. Related work

Misogyny detection is closely related to hate speech and abusive language detection, but it introduces specific challenges around gendered stereotypes, sexual objectification, symbolic violence, and context-dependent interpretation [5]. In song lyrics, these challenges are amplified by figurative language, narrative voice, repetition, and genre conventions: a lyric may be aggressive or sexualized without being misogynistic, while another may encode misogyny through indirect possessiveness or stereotype rather than an explicit insult [1].

The closest context for MiSonGyny 2026 is a line of IberLEF shared tasks on hate speech, sexism, and misogyny. MiSonGyny 2026 was organized within IberLEF 2026 [2] and builds directly on MiSonGyny 2025, which introduced misogyny detection in Spanish-language song lyrics as a shared-task setting [3]. A closely related predecessor is HOMO-MEX 2024, whose third subtask addressed hate speech detection in Spanish-language song lyrics targeting the Mexican Spanish-speaking LGBT+ population [4]. Earlier Iberian evaluation campaigns addressed related phenomena in social-media text: AMI at IberEval 2018 focused on automatic misogyny identification in Spanish and English tweets, including misogyny categorization and target classification [6], while EXIST at IberLEF 2021 focused on sexism identification in social networks, including binary and fine-grained classification in Spanish and English [7]. MiSonGyny 2026 differs from these earlier settings by combining misogyny detection with long, stylistically dense song lyrics and a hierarchical task structure.

From a modeling perspective, these tasks are commonly approached as supervised text-classification problems. Transformer encoders are widely used as backbones for text classification [10, 11]. Multilingual models such as XLM-RoBERTa [8] are attractive for Iberian and Latin American Spanish because they provide strong pretrained representations without task-specific pretraining.

Lyrics, however, routinely exceed the input length of standard encoders, requiring truncation, chunking, or hierarchical aggregation. Our system adopts a pragmatic first/last token-selection strategy rather than a heavier long-context architecture. We further rely on two well-established shared-task strategies: multi-task learning, which shares representations across related subtasks while task-specific heads capture distinct decision boundaries, and threshold tuning, which is especially important under macro-F_1 and class imbalance, where the default 0.5 rule is often poorly aligned with the metric.

2.2. Subtasks and hierarchical structure

MiSonGyny 2026 defines three subtasks, all evaluated independently using macro-F_1:

- **T1 – Misogyny Detection.** Binary song-level classification between Misogynistic (M) and Non-Misogynistic (NM).
- **T2 – Misogyny Type Classification.** Multilabel categorization defined only for misogynistic songs, with three possible labels: Sexualization (S), Violence (V), and Hate (H).

- **T3 – Stereotype Identification.** Binary classification of whether a misogynistic song contains gender stereotypes (Y/N).

Inspection of the released training data confirms a strict hierarchical dependency between tasks. Songs annotated as non-misogynistic never receive type labels (Sexualization, Violence, or Hate) and are always assigned a negative stereotype label. In practice, this means that T2 and T3 are only defined for songs classified as misogynistic. Following this convention, our inference pipeline treats T1 as a deterministic gate over the downstream tasks (Section 3.4).

2.3. Dataset statistics

Table 1 summarizes the corpus. The dataset contains 2,303 training songs and 576 test songs, with 43.8% of the training examples labeled as misogynistic. Conditional on misogyny, Sexualization is the most frequent type label (79.7%), followed by Hate (58.4%) and Violence (31.1%). Lyric length is the dominant practical constraint. The median song contains 308 words, the 95th percentile reaches 714 words, and the longest lyric contains 2,258 words. We estimate that approximately 33.8% of songs exceed the 512-token input budget of XLM-RoBERTa under standard subword tokenization, motivating the token-selection strategy described in Section 3. The corpus is Spanish with mixed regional varieties (es_ES, es_LATAM); urban-music slang, code-switching, and onomatopoeia are consistent across splits and recur in the error patterns of Section 5.

Table 1
MiSonGyny 2026 corpus statistics. T2 and T3 priors are conditional on M.

Quantity	Value	Notes
Training songs	2,303	
Test songs	576	
M / NM (train)	43.8% / 56.2%	1,009 / 1,294
T2 prior S	79.7%	majority
T2 prior H	58.4%	intermediate
T2 prior V	31.1%	minority class
T3 prior Y / N	68.1% / 31.9%	
Median / p95 / max len	308 / 714 / 2,258	words
$>$512 tokens	$\sim$33.8%	motivates token-selection strategy

3. System Overview

Our submitted system is a gated multi-task model built on top of XLM-RoBERTa-large. The model receives a song lyric as input, reduces it to a maximum-length subword sequence, encodes it once with a shared transformer encoder, and predicts the three MiSonGyny tasks through parallel task-specific heads. A deterministic inference-time gate is then applied to enforce the annotation hierarchy of the dataset: when a song is predicted as non-misogynistic in T1, the downstream T2 and T3 outputs are suppressed. Figure 1 shows the full pipeline: a lyric is reduced by first/last token selection, encoded by XLM-RoBERTa-large and mean-pooled, projected by three task heads, and resolved at inference by a hierarchical T1 gate.

3.1. Input encoding and shared representation

Lyrics were tokenized using the SentencePiece tokenizer associated with XLM-RoBERTa-large. We selected this encoder because it provides pretrained multilingual representations that include Spanish [8]. Since many lyrics exceed the encoder input limit, we used first/last token selection. For over-length lyrics, we retain 255 subword tokens from the beginning and 255 from the end, concatenate them into

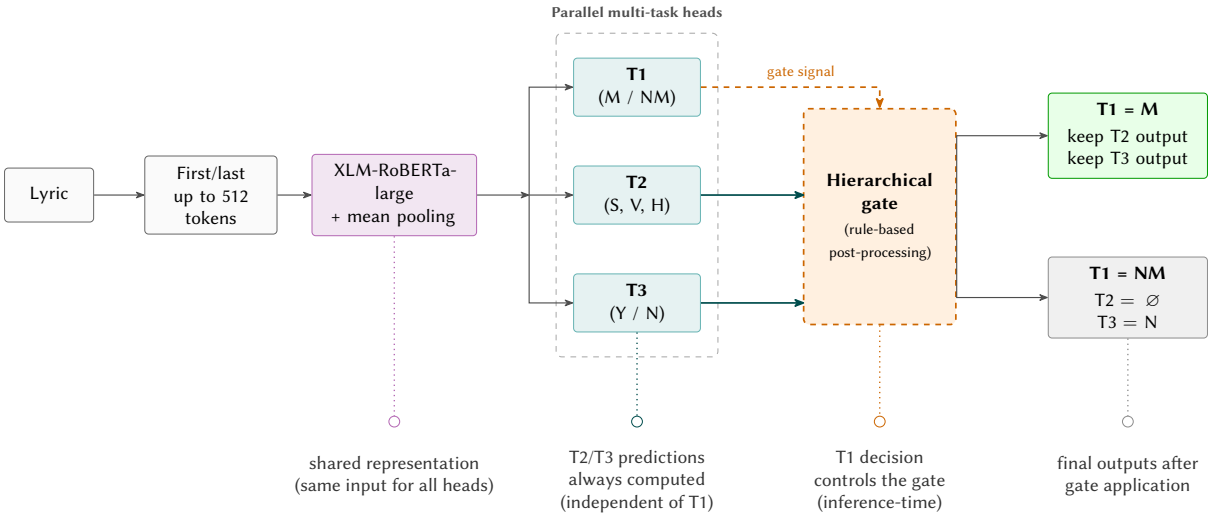

Figure 1: System pipeline. A song lyric is reduced with a first/last token-selection strategy and encoded once by XLM-RoBERTa-large with mean pooling. The shared representation feeds three *parallel* task heads (T1, T2, T3), all computed independently. The *hierarchical gate* is not a neural component but a deterministic, rule-based inference-time post-processing step: it reads the T1 decision and either keeps the T2/T3 predictions (T1=M) or suppresses them (T1=NM $\Rightarrow$ T2=$\varnothing$, T3=N).

a 510-token content sequence, and add the two special boundary tokens required by XLM-RoBERTa, yielding 512 tokens in total. Shorter lyrics are encoded in full and padded. This strategy preserves information from both the opening and closing parts of the lyric while reducing over-length inputs to the encoder limit.

The resulting sequence is encoded with XLM-RoBERTa-large. We obtain a single lyric-level representation by applying masked mean pooling over all non-padding positions of the final-layer hidden states, including the special boundary tokens. This produces one 1024-dimensional vector per lyric. The same pooled vector is used as the input to all task-specific heads.

3.2. Multi-task prediction heads

The model has three task-specific linear heads, all fed by the same shared representation. T1 is a binary misogyny detection head that maps the 1024-dimensional vector to one logit and estimates whether the lyric is misogynistic (M) or non-misogynistic (NM). T2 is a multilabel head that maps the same vector to three independent logits, corresponding to Sexualization (S), Violence (V), and Hate (H). Since these labels are not mutually exclusive, sigmoid activations are applied independently to the three outputs. T3 is a binary stereotype identification head that maps the shared vector to one logit and predicts whether the lyric contains a stereotype (Y/N). A shared dropout layer with probability 0.1 is applied to the pooled representation before it is passed to the three task heads.

The three heads are evaluated in parallel. T2 and T3 do not receive the T1 prediction as input, and no head consumes the logits, probabilities, or decisions produced by another head. The hierarchy is therefore not encoded as a neural cascade; it is enforced only after prediction through a deterministic inference-time gate described below.

3.3. Training, model selection, and full retraining

The model was trained jointly using the unweighted sum of the three task losses. All heads use Focal Loss with a focusing parameter of 2.0. For T2, we use positive-class weights [0.254, 2.215, 0.712] for [S, V, H], respectively, derived from the label frequencies of misogynistic training songs. Because T2 and T3 labels are defined only for lyrics annotated as misogynistic, their losses are computed only for training instances whose gold T1 label is M. T1 is trained on all instances, while the T2 and T3 losses

are masked for non-misogynistic examples. Thus, the model learns all three tasks jointly, but the T2 and T3 losses are applied only to misogynistic training instances.

We trained the model on an internal 80/20 train-development split and selected the checkpoint with the highest average development F1 across T1, T2, and T3. Threshold optimization was performed only after checkpoint selection. From this checkpoint, we computed development-set probabilities and selected one threshold per binary decision by five-fold cross-validation. Final thresholds were obtained as the mean of the fold-wise optima. Finally, the selected checkpoint was used to initialize a three-epoch full-retraining stage over the full training set. Learning rates for both the encoder and the task heads were scaled by a factor of 0.4 during this stage. The three-epoch duration and the 0.4 learning-rate scaling factor were set heuristically; neither was selected through a separate hyperparameter search or compared against alternative full-retraining configurations.

3.4. Thresholding and hierarchical inference

At inference time, the three heads produce probabilities for every lyric. The submitted system used task-specific thresholds selected on the development split: 0.680 for T1; 0.356, 0.240, and 0.204 for the three T2 labels (S, V, H); and 0.090 for T3.

After thresholding, we apply a deterministic hierarchical gate. If T1 predicts M, the T2 and T3 predictions are kept unchanged. If T1 predicts NM, the system emits no positive type labels and forces T3 to N. Formally, for any lyric predicted as non-misogynistic,

$$\mathrm{T1} = \mathrm{NM} \Rightarrow (\mathrm{S} = 0 \wedge \mathrm{V} = 0 \wedge \mathrm{H} = 0 \wedge \mathrm{T3} = \mathrm{N})$$

This gate is a rule-based post-processing step and has no trainable parameters.

For lyrics predicted as misogynistic, T2 is allowed to emit any subset of the three type labels. If thresholding produces no positive T2 label for a lyric predicted as M, the system assigns the type label with the highest T2 probability. This avoids empty type outputs for lyrics that the system has already classified as misogynistic. The final submitted labels are therefore obtained by combining task probabilities, development-selected thresholds, and the deterministic hierarchical gate.

4. Results and Analysis

4.1. Official Results

The official Codabench macro-F1 scores of the submitted system and the corresponding development scores are summarized in Table 2.

Table 2
Official MiSonGyny 2026 results for aerojash. Development scores correspond to the internal 20% held-out split; $\Delta = \mathrm{Test} - \mathrm{Devel}$.

Task	Devel	Test	Δ	Rank
T1 (M/NM)	0.8149	0.8581	$+0.043$	3rd / 11
T2 (S/V/H)	0.7518	0.6675	-0.084	–
T3 (Y/N stereotype)	0.7052	0.8071	$+0.102$	3rd / 11

T1 and T3 achieved third-place rankings, whereas T2 obtained a lower official performance despite stronger development results. The asymmetry between development and test performance motivates a closer analysis of the system's behavior across tasks. For T1 and T3, the official test scores use the subsequent full-retrain model; therefore, Δ should be interpreted as a validation-to-submission discrepancy rather than as a pure generalization gap.

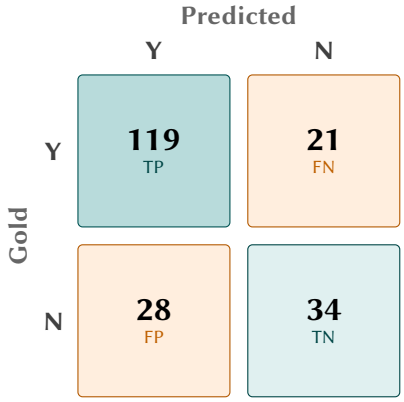

Figure 2: T3 development confusion matrix (stereotype detection, Y / N). Diagonal cells (TP / TN) are shaded in teal; off-diagonal errors (FN / FP) are shown in muted amber.

4.2. Development–test asymmetry

The same fine-tuning procedure produced opposite-direction development–test differences across tasks (Table 2). Three non-mutually-exclusive factors are involved, none isolated by our design. *(a) Full-retrain.* Both T1 and T3 were generated with the full-retrained model after three additional epochs on the full corpus; the differences ($+0.043$, $+0.102$) are consistent with a benefit from this step but are not separated from threshold and distributional effects. *(b) Threshold application.* The T2 score decrease coincides with per-label thresholds that transferred less effectively from development to the full-retrained model than the corresponding T1 and T3 thresholds. *(c) Distributional shift.* The T3 operating point behaves differently on development and test, suggesting a combination of calibration and distributional effects explored in Section 4.3. Taken together, these observations suggest that the development–test differences arise from multiple interacting factors rather than from a single failure mode.

4.3. Threshold Behavior Across Development and Test

The submitted T3 threshold of 0.090 defines an unusually permissive operating point. On the official test set, the system predicts Y for approximately 94% of songs predicted as misogynistic. We retained this threshold because the resulting development macro-F1 was 0.7052, and similarly low operating points were selected consistently across folds.

The selected threshold was therefore low but not degenerate on development. Figure 2 shows that both positive and negative classes remain distinguishable at the submitted operating point, yielding per-class F1 scores of 0.83 for Y and 0.58 for N.

The more notable observation is the gap between the development prediction rate and the 94% test prediction rate. Under the same threshold, the system produces substantially more positive T3 predictions on the test than on the development. Two possibilities are consistent with our data: a higher Y proportion on the test or a shift in the distribution of model scores produced by the v1 full-retrain model on the test. Because the official test labels are unavailable, neither possibility can be established. We therefore report this threshold as an observed operating point in our run, not as evidence of a general property of macro-F1 under imbalance.

4.4. Ensemble Behavior

We built a three-source ensemble combining an XGBoost [12] frozen-embedding baseline and two independently fine-tuned XLM-RoBERTa models. Ensemble weights were obtained through a development-set grid search maximizing T1 macro-F1. The ensemble produced a small improvement on development but erased the test advantage of the submitted full-retrain model (Table 3).

Table 3 shows that the ensemble improved T1 from 0.8149 to 0.8217 on development ($+0.007$) but decreased the official test score from 0.8581 to 0.8132 (-0.045). Thus, the direction of the development gain reversed on the test set.

Table 3
Ensemble vs. single full-retrain model on T1 (macro-F_1). The development gain reverses on test.

System	Devel	Test
full-retrain (single)	0.8149	**0.8581**
Ensemble (v1+v2+XGB)	**0.8217**	0.8132
Δ	$+0.007$	-0.045

Several explanations are consistent with the observed reversal. One possibility is that the development-set weight search did not adequately reflect the test-time behavior of the submitted full-retrain model, causing the ensemble to dilute the contribution of its strongest component. Other possibilities include weight-search overfitting on the relatively small development split, mismatched probability scales across uncalibrated sources, seed variance, test-composition idiosyncrasy, or the fixed XGBoost weight. Because our experiments were not designed to isolate these factors, we cannot distinguish among them.

By contrast, the same averaging strategy did not harm T2: a two-model ensemble of the fine-tuned encoders (without the XGBoost source) yielded a marginal gain on the official test set, from 0.6675 to 0.6699. We evaluated this task-specific configuration as a separate, post-hoc submission rather than as part of the primary system reported in Table 2. The contrast between T1 (regression) and T2 (marginal gain) under the same averaging strategy is therefore an empirical observation rather than a principled methodological contribution.

5. Error Analysis

We inspected misclassifications on the held-out development split (461 songs: 259 NM and 202 M) using the submitted thresholds. Overall, 258 songs (56.0%) accumulated errors in two or more head-level subdecisions, and 10 songs (2.2%) accumulated errors in three or more. These counts are computed from per-head predictions before applying the deterministic hierarchical gate. Therefore, they should be interpreted as a diagnostic measure of head-level disagreement with the gold labels rather than as the exact number of errors retained in the final submitted outputs. For songs predicted as non-misogynistic by T1, the gate suppresses the corresponding T2 and T3 outputs.

For T1, false positives were frequently associated with urban-music lyrics rich in slang, sexual innuendo, and forms of direct address such as “mami” or “baby”. These songs received high misogyny probabilities despite non-misogynistic gold labels. In contrast, false negatives were often slower ballads expressing possessiveness, emotional manipulation, or objectification through metaphorical language, for example “fuiste mía, sólo mía”. Compared with the false positives, these cases tended to rely more heavily on implicit meaning than on explicit lexical cues.

The most difficult T2 cases involved the Violence label. False positives frequently corresponded to non-gender-targeted aggression, such as gangster imagery, drinking narratives, or aggressive expressions directed at unspecified targets. False negatives, in contrast, often depict violence indirectly through metaphor or implication. These observations suggest that distinguishing misogynistic violence from more general forms of aggression remained challenging for the model.

A similar ambiguity appeared in T3. False negatives were often short or opaque lyrics in which stereotypes were expressed through brief recurring tropes, whereas false positives tended to be longer songs whose repeated thematic patterns were interpreted by the model as stereotypical despite a negative annotation. As with T1 and Violence, these disagreements were frequently associated with indirect or context-dependent expressions rather than explicit statements.

Taken together, these observations suggest that the most difficult songs were boundary cases rather than examples characterized by a single source of error. Under the hierarchical inference scheme, such cases are particularly important because errors in T1 can propagate to downstream decisions. Our analysis does not disentangle the contribution of the hierarchy from the intrinsic difficulty of the examples, but it identifies the types of lyrics for which the submitted system remained most vulnerable.

6. Limitations

The findings above rest on a constrained experimental setup, and several caveats apply throughout. First, all results come from a single training corpus and a single 80/20 development split with no repeated seeds, so the stability of the reported patterns is unverified. Second, the first/last token selection is not compared against alternative approaches for long-document processing, and its contribution cannot be isolated from that of the underlying encoder. Third, we did not submit a matched version of the model without the full-retrain stage, so the effects of full retraining, threshold selection, and distributional differences cannot be cleanly separated. Finally, the official test labels are unavailable, so all explanations involving test-set behavior remain indirect and are inferred from leaderboard results and submission outputs. Some observed disagreements may additionally reflect annotation uncertainty rather than model deficiency; multi-annotator labels would help disentangle these sources.

7. Conclusions

The aerojash submission to MiSonGyny 2026 achieved macro-F1 scores of 0.8581 on T1 (3rd/11), 0.8071 on T3 (3rd/11), and 0.6675 on T2 using a hierarchical multi-task architecture based on XLM-RoBERTa-large, first/last token selection, task-specific thresholds, and a deterministic inference gate.

Beyond the official rankings, the competition provided an opportunity to examine how a hierarchical classifier behaves under realistic shared-task conditions. Development and test performance evolved differently across tasks; the submitted T3 threshold produced an unusually permissive operating point, and a development-selected ensemble failed to retain its advantage on the official test set. Although our experiments do not isolate the causes of these effects, they illustrate how model selection, thresholding decisions, and hierarchical inference can influence observed performance.

Qualitative inspection further suggests that the most difficult songs are those in which misogyny, violence, and stereotypes are expressed indirectly or ambiguously. Under a hierarchical prediction scheme, such boundary cases are especially important because errors in the top-level decision can affect downstream outputs.

We hope that both the system description and the accompanying analyses are useful to future participants in MiSonGyny and related shared tasks involving long, culturally situated texts.

Acknowledgments

The authors thank the MiSonGyny 2026 organizers for the shared task design and the Codabench evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to check grammar and spelling, improve writing style, assist with formatting, simulate peer review, support debugging, and inspect author-written code. The authors reviewed and edited the outputs as needed; they take full responsibility for the publication's content.

References

[1] T. Alcántara, E. Urios Alacreus, C. Macías, M. Aloy Mayo, O. García-Vázquez, A. Espinosa-Juárez, M. Soto, P. Rosso, J. E. Valdez-Rodríguez, and H. Calvo, *Overview of MiSonGyny at IberLEF 2026: Misogyny Speech Detection in Mexican and Spanish Language Song Lyrics*, Procesamiento del Lenguaje Natural (2026).

[2] A. Bonet-Jover, J. A. González-Barba, and 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.

[3] T. Alcántara, M. Soto, C. Macías, O. García-Vázquez, A. Espinosa-Juárez, H. Calvo, J. E. Valdez-Rodríguez, and E. Felipe-Riverón, *Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics*, Procesamiento del Lenguaje Natural 75 (2025).

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

[5] A. Dutta, S. Banducci, and C. Q. Camargo, *Divided by discipline? A systematic literature review on the quantification of online sexism and misogyny using a semi-automated approach*, Scientometrics 130(9) (2025) 4915–4971. doi:10.1007/s11192-025-05410-2.

[6] E. Fersini, P. Rosso, and 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), CEUR Workshop Proceedings, vol. 2150, 2018, pp. 214–228.

[7] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso, *Overview of EXIST 2021: sEXism Identification in Social neTworks*, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, *Unsupervised Cross-Lingual Representation Learning at Scale*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal Loss for Dense Object Detection*, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in: Advances in Neural Information Processing Systems 30 (NeurIPS), Curran Associates, 2017, pp. 5998–6008.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, arXiv preprint arXiv:1907.11692, 2019.

[12] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.