CICESE-LDCAA at MiSonGyny 2026: Detection of Misogyny in Song Lyrics Using Machine Learning

Juliette Ninnie Lazo-Vazquez[1,2,*], Joan Manuel Raygoza-Romero[1,2], Jesus Antonio Navarrete-Lopez[1], Cielo Aholiva Higuera-Gutiérrez[1] and Irvin Hussein López-Nava[1,2]

[1]*Data Science and Machine Learning Lab, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Tijuana-Ensenada Highway 3918, Ensenada, BC 22860, Mexico*

[2]*Facultad de Ciencias, Universidad Autónoma de Baja California (UABC), Transpeninsular Highway 3917, Ensenada, BC 22860, Mexico*

Abstract

This paper describes the participation of the CICESE-LDCAA team in the MiSonGyny 2026 shared task, focused on the detection and characterization of misogynistic content in Spanish language song lyrics. The system addresses three subtasks: binary misogyny detection, multilabel misogyny type identification, and gender stereotype detection. The proposed methodology follows a hierarchical strategy in which a Longformer-based classifier first determines whether a lyric is misogynistic. When misogyny is detected, ensemble classifiers supported by Qwen 3 embeddings and RoBERTuito-derived features are used to identify sexualization, violence, hate, and gender stereotypes. Official results show competitive performance across the three subtasks, with F1-scores of 0.8533 in Task 1, 0.6846 in Task 2, and 0.7924 in Task 3, ranking 5th, 4th, and 6th place, respectively. These results suggest that combining long-context transformers, semantic representations, linguistic features, and ensemble learning is a promising approach for analyzing misogyny and gender stereotypes in song lyrics.

Keywords

Song lyrics, Natural Language Processing, Machine learning, Transformers, Longformer, Qwen embeddings, RoBERTuito, Ensemble learning

1. Introduction

Misogyny is a deeply rooted social problem that affects women across different cultural, economic, and communicative contexts. The MiSonGyny 2026 shared task was organized as part of IberLEF 2026, a forum devoted to Natural Language Processing challenges for Spanish and other Iberian languages [1]. In MiSonGyny 2026, misogyny is understood as any form of discrimination, prejudice, or aversion toward women based on their gender. Its consequences include gender-based violence, economic inequalities, social exclusion, mental health issues, and unequal access to education and job opportunities [2]. From the perspective of Natural Language Processing (NLP), this phenomenon is especially relevant because misogynistic discourse can be expressed not only through explicit hate, but also through subtle, symbolic, or normalized forms of language.

Music represents one of the most influential forms of cultural expression. Song lyrics can reflect social beliefs, emotions, and identities, but they can also reproduce harmful stereotypes and discriminatory attitudes. In this context, the MiSonGyny 2026 overview emphasizes that music often perpetuates negative stereotypes about women through lyrics that objectify, sexualize, or demean them, contributing to the normalization of misogynistic attitudes [2]. Previous large-scale studies on song lyrics have also

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

juliette.lazo@uabc.edu.mx (J. N. Lazo-Vazquez); jraygoza@cicese.edu.mx (J. M. Raygoza-Romero); jnavarrete@cicese.edu.mx (J. A. Navarrete-Lopez); aholiva@cicese.edu.mx (C. A. Higuera-Gutiérrez); hussein@cicese.edu.mx (I. H. López-Nava)

https://github.com/JulietteLazo (J. N. Lazo-Vazquez)

0009-0002-6865-4468 (J. N. Lazo-Vazquez); 0000-0003-3085-5678 (J. M. Raygoza-Romero); 0009-0009-2890-4822 (J. A. Navarrete-Lopez); 0009-0005-1172-2679 (C. A. Higuera-Gutiérrez); 0000-0003-3979-9465 (I. H. López-Nava)

shown that musical texts may contain measurable patterns of sexism and gender bias, confirming the relevance of lyrics as a source for computational analysis of gender-related discourse [3].

Automatically detecting misogyny in song lyrics is a challenging NLP problem. Unlike other types of text, lyrics frequently include figurative language, repetitions, slang, implicit meanings, cultural references, and ambiguous expressions. These characteristics make it difficult for computational models to determine whether a lyric contains misogynistic content, what type of misogyny is present, or whether it reinforces gender-based stereotypes. Moreover, misogynistic discourse can overlap with broader forms of hate speech, which the United Nations defines as offensive discourse targeting individuals or groups based on inherent characteristics such as gender, religion, or race [4]. Therefore, the task requires models capable of capturing both explicit and implicit forms of harmful language.

The MiSonGyny 2026 shared task addresses this problem by focusing on the detection and analysis of misogyny in Mexican and Spanish language song lyrics [2]. The challenge is organized into three complementary subtasks: binary misogyny detection, multilabel misogyny type classification, and gender stereotype identification. Task 1 aims to determine whether a song lyric contains misogynistic discourse. Task 2 identifies the specific types of misogyny present in the lyric, considering sexualization, violence, and hate as non-exclusive categories. Task 3 focuses on detecting gender-based stereotypes, allowing the analysis to capture biased or restrictive representations of gender even when the lyric is not explicitly violent or hateful.

This work is also related to previous shared tasks on harmful language detection in Spanish. MiSonGyny 2025 addressed misogyny speech detection in Spanish song lyrics [5], while HOMO-MEX 2024 focused on hate speech detection toward the Mexican Spanish-speaking LGBT+ population [6]. These shared tasks show the increasing relevance of computational approaches for analyzing abusive, discriminatory, and gender-related discourse in Spanish-language contexts.

In this work, we describe the participation of the CICESE-LDCAA team in the MiSonGyny 2026 challenge. Our approach explores different NLP and machine learning strategies for Spanish-language song lyrics, including classical machine learning models, TF-IDF representations, preprocessing variants, and transformer-based models. Special attention is given to the main challenges of the task: class imbalance, multilabel prediction, long and noisy musical texts, and the subtle nature of misogynistic and stereotypical discourse. Through this study, we aim to evaluate the performance of different modeling approaches for detecting and characterizing harmful gender-related content in song lyrics.

2. Task and Dataset Description

The MiSonGyny 2026 shared task focuses on the automatic detection and characterization of misogynistic content in Spanish language song lyrics [2]. The task is organized into three subtasks that address different levels of analysis: binary misogyny detection, misogyny type identification, and gender stereotype detection. These subtasks are based on the same textual input and allow each lyric to be analyzed from complementary perspectives.

Unlike previous editions, the 2026 edition provides a unified dataset in a single CSV file. This file contains the information required for the three subtasks: misogyny detection, misogyny type identification, and gender stereotype detection. Therefore, each song lyric can be analyzed from multiple perspectives using the same textual input.

2.1. Dataset Description

The training dataset contains 2,303 instances of Spanish song lyrics. Each instance corresponds to a song and includes metadata about the song and artist, the full lyrics, the language variant, and the annotation labels associated with the three subtasks. The dataset includes lyrics from two Spanish variants: Latin American Spanish (`es_LATAM`) and Peninsular Spanish (`es_ES`). Table 1 summarizes the columns included in the dataset.

Table 1
Description of the columns included in the MiSonGyny 2026 training dataset.

Column	Description
song_id	Unique identifier of the song.
song_title	Title of the song.
artist_id	Unique identifier of the artist.
artist_name	Name of the artist.
lyrics	Full lyrics of the song.
language	Spanish language variant of the lyric.
is_misogynistic	Binary label for misogyny detection.
type_sexualization	Binary label for sexualization.
type_violence	Binary label for violence.
type_hate	Binary label for hate.
has_gender_stereotype	Binary label for gender stereotype detection.

2.2. Task 1: Misogyny Detection

Task 1 is a binary classification problem. The objective is to determine whether a song lyric contains misogynistic content. The target column for this task is `is_misogynistic`, which contains two possible labels:

- **Misogynistic (M):** The lyric contains misogynistic content, such as discrimination, contempt, degradation, objectification, sexualization, hostility, or harmful stereotypes directed toward women.
- **Not Misogynistic (NM):** The lyric does not contain misogynistic content. It may mention women, relationships, or gender-related topics, but it does not reinforce misogynistic attitudes or harmful stereotypes.

The class distribution for Task 1 is presented in Table 2. The dataset shows a slight imbalance toward the non-misogynistic class, although both categories are sufficiently represented for binary classification.

Table 2
Class distribution for Task 1: Misogyny Detection.

Class	Number of instances	Percentage
Misogynistic (M)	1,009	43.8%
Not Misogynistic (NM)	1,294	56.2%
Total	**2,303**	**100.0%**

2.3. Task 2: Misogyny Type Identification

Task 2 is a multilabel classification problem. The goal is to identify the specific types of misogynistic content present in each lyric. This task is represented by three binary columns: `type_sexualization`, `type_violence`, and `type_hate`.

In contrast to a multiclass setting, the misogyny-related categories considered in Task 2 are not mutually exclusive. A single lyric may express multiple forms of misogynistic content simultaneously, such as sexualization combined with hate, or violence with sexualization. Consequently, Task 2 is formulated as a multilabel classification problem, in which the model must independently predict the presence or absence of each label for every song.

The three labels considered in this task are defined as follows:

- **Sexualization:** Lyrics that include sexual objectification, sexual references, insinuations, or expressions that reduce women to their bodies or sexual availability.

- **Violence:** Lyrics that include physical or verbal aggression, threats, coercion, domination, or violent actions directed toward women.
- **Hate:** Lyrics that contain offensive, hostile, demeaning, or discriminatory language against women or groups of women.

In the dataset, the type labels are annotated for the misogynistic subset. Non-misogynistic instances contain missing values in these columns because no misogyny type is applicable. Table 3 shows the number of positive and negative labels for each misogyny type among the misogynistic lyrics.

Table 3
Label distribution for Task 2: Misogyny Type Identification.

Misogyny type	Positive labels	Negative labels	Positive percentage
Sexualization	804	205	79.68%
Violence	314	695	31.12%
Hate	589	420	58.37%

This distribution shows that sexualization is the most frequent misogyny type, followed by hate and violence. The lower frequency of violence indicates a class imbalance that may affect model performance, especially when optimizing macro-averaged metrics.

2.4. Task 3: Gender Stereotype Detection

Task 3 is a binary classification problem focused on identifying whether a misogynistic lyric contains gender stereotypes. The target column for this task is `has_gender_stereotype`, with two possible labels:

- **Yes (Y):** The lyric contains gender stereotypes.
- **No (N):** The lyric does not contain identifiable gender stereotypes.

A lyric is considered to contain a gender stereotype when it reinforces traditional, restrictive, or harmful assumptions about women or men. These stereotypes may be related to appearance, behavior, emotional roles, romantic expectations, dependency, submission, dominance, or socially imposed gender norms. As in Task 2, this label is annotated for the misogynistic subset of the dataset. Non-misogynistic lyrics contain missing values for this column. Table 4 shows the distribution of labels for Task 3.

Table 4
Class distribution for Task 3: Gender Stereotype Detection.

Class	Number of instances	Percentage
Gender stereotype present (Y)	687	68.09%
Gender stereotype absent (N)	322	31.91%
Total	**1,009**	**100.00%**

This task complements the previous subtasks because gender stereotypes may appear even when the lyric is not explicitly violent or hateful. Therefore, Task 3 allows the analysis to capture more implicit forms of gender bias in music.

3. Methodology

The proposed methodology followed a hierarchical prediction strategy for the three subtasks of the MiSonGyny 2026 shared task, as illustrated in Figure 1. First, Task 1 was addressed as a binary misogyny detection problem. The prediction obtained in this stage was then used as a decision gate for Task 2 and Task 3. If a lyric was classified as non-misogynistic, the outputs for the remaining tasks are directly

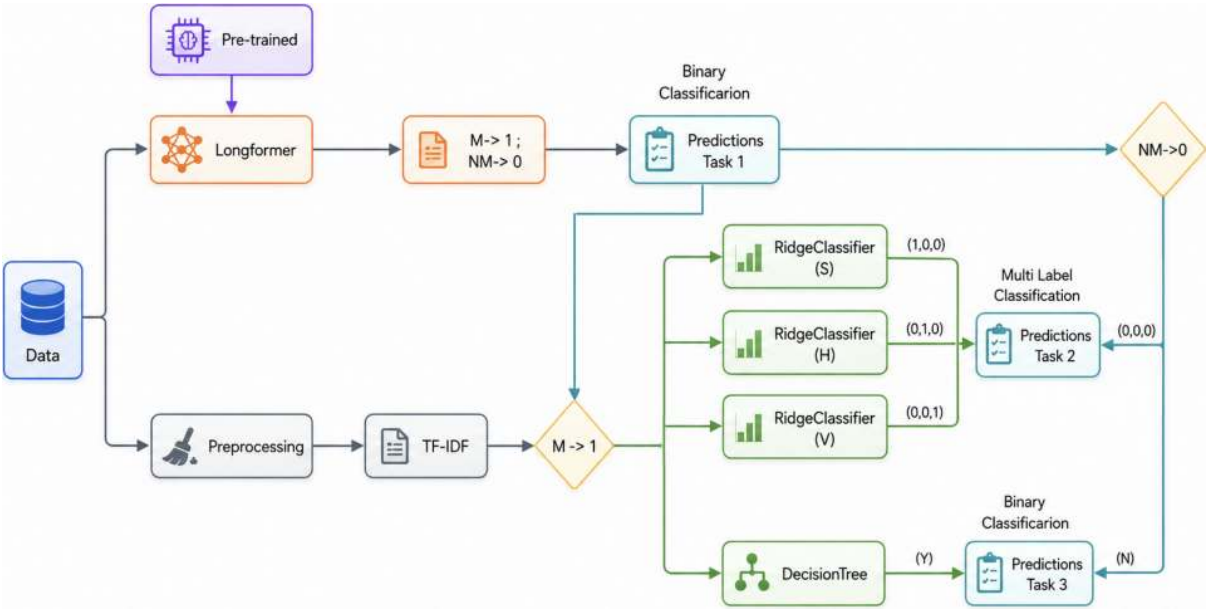

Figure 1: Hierarchical methodology followed for the three tasks. Task 1 is first addressed with a Longformer binary classifier. Its prediction is then used as a decision gate for Task 2 and Task 3: non-misogynistic lyrics are assigned negative labels, while misogynistic lyrics are processed by the Task 2 and Task 3 classifiers.

assigned as negative labels: $(0, 0, 0)$ for Task 2 and N for Task 3. If the lyric is classified as misogynistic, the system activated the classifiers for misogyny type identification and gender stereotype detection.

This hierarchical design was adopted because the labels of Task 2 and Task 3 are directly related to the presence of misogynistic content. Therefore, Task 1 provides the first decision layer of the system, while Task 2 and Task 3 focus on finer-grained distinctions within misogynistic lyrics, such as identifying sexualization, violence, hate, and gender stereotypes.

3.1. Data Partitioning

For the experimental setup, the dataset was organized into training, validation, and internal test partitions following an approximate 70/20/10 distribution, respectively. The split was performed in a stratified manner in order to preserve the class distribution of the corresponding task. The validation partition was mainly used for transformer-based models, where a fixed validation set was required during training. For the remaining machine learning models, model selection and performance estimation were carried out using cross-validation. In all cases, the internal test set was kept independent and was not used during training, hyperparameter optimization, feature selection, or threshold tuning.

3.2. TF-IDF-Based Machine Learning Models

As an initial modeling strategy, classical machine learning models were evaluated using TF-IDF representations. These experiments were designed to provide competitive baselines and to analyze the effect of different preprocessing configurations on classification performance.

A total of 32 preprocessing variants were generated from the combination of five optional techniques: stopword removal, lemmatization, cleaning, stemming, and POS-based token selection. The cleaning process included lowercasing, Unicode normalization, removal of URLs, user mentions, hashtag symbols, numbers, and non-informative punctuation. The POS-based token selection step kept informative part-of-speech categories such as nouns, proper nouns, verbs, auxiliaries, adjectives, and adverbs.

After preprocessing, the lyrics were represented using TF-IDF features. The vectorizer used word unigrams and bigrams, sublinear term frequency scaling, and document frequency filtering. For most

experiments, the configuration included `min_df=2`, `max_df=0.98`, and a maximum of 30,000 features.

The evaluated classifiers included Logistic Regression, Linear SVM, SGDClassifier, RidgeClassifier, Passive Aggressive, Perceptron, Multinomial Naive Bayes, Complement Naive Bayes, Decision Tree, Random Forest, XGBoost, CatBoost, AdaBoost, and LightGBM. These models were implemented using Scikit-learn [7] and related machine learning libraries.

For Task 1, these models were used as binary classifiers for misogyny detection. For Task 2, the multilabel problem was decomposed into three binary classification problems using a binary relevance strategy, with one classifier for each label: sexualization, violence, and hate. For Task 3, the models were evaluated as binary classifiers for gender stereotype identification.

3.3. Transformer-Based Models

In addition to classical machine learning models, transformer-based models were explored for text classification. These experiments were implemented using the Hugging Face Transformers library [8] with PyTorch [9], considering pretrained architectures related to BERT [10], RoBERTa [11], and Longformer [12].

The final transformer-based model used for Task 1 was Longformer. This architecture was selected because song lyrics may contain long textual sequences, repetitions, and contextual dependencies that can be difficult to capture with standard transformer models that have shorter input limits. The Longformer model was initialized from the Hugging Face checkpoint `joheras/longformer-base-4096-bne-es-finetuned-clinais` and fine-tuned for single-label binary classification. Although this checkpoint was originally fine-tuned on clinical Spanish text, it was selected because it provides a Spanish long-context initialization suitable for processing long lyrics. We acknowledge that this domain mismatch may limit its ability to model musical language, slang, and figurative expressions.

Before training, rows with missing lyrics, missing labels, or empty text were removed. The labels were normalized and encoded as numerical values, where *NM* was mapped to 0 and *M* was mapped to 1. The lyrics were converted into Hugging Face datasets and tokenized with the corresponding Longformer tokenizer. A maximum sequence length of 2048 tokens was used. The tokenized dataset contained `input_ids`, `attention_mask`, and the corresponding label.

Training was performed for five epochs with a learning rate of 2×10^{-5}, weight decay of 0.01, warmup steps, gradient accumulation, and early stopping. The best checkpoint was selected according to the validation F1-score. The final predictions produced by this Longformer model were used as the main decision gate for Task 2 and Task 3.

3.4. Embedding-Based Representation with Qwen and RoBERTuito Features

Given the limitations of sparse lexical representations such as TF-IDF in modeling semantic and contextual dependencies, additional feature representations were used into the proposed methodology. This is particularly relevant for song lyrics, whose meaning may be expressed through figurative language, slang, implicit references, and repeated patterns. To enrich textual representations, two complementary information sources were considered: contextual embeddings and features derived from pretrained models targeting different linguistic phenomena.

First, Qwen was used to extract contextual embeddings from the complete lyric. Unlike TF-IDF, these embeddings provided dense semantic representations that can capture contextual relationships between words and phrases. This is especially useful for song lyrics, where meaning can depend on figurative language, implicit references, slang, repetitions, and broader textual context.

For each lyric, embeddings were extracted using `Qwen/Qwen3-8B`, a 8-billion-parameter model from the Qwen3 family available on Hugging Face [13]. Each lyric was represented as a 4096-dimensional embedding vector, which was used as one of the main feature sources for the final classifiers.

Qwen3 was used as a frozen feature extractor to leverage its pretrained semantic knowledge while maintaining a computationally efficient and reproducible pipeline. Given the limited size of the shared-

task dataset, full fine-tuning of a 8B-parameter model could increase the risk of overfitting and introduce substantially higher training costs. In addition to Qwen embeddings, complementary features were extracted using RoBERTuito-based models specialized in sentiment analysis, emotion recognition, hate speech detection, and irony identification [14].

Since song lyrics often exceed the model input length, a sliding-window strategy was employed, dividing each lyric into segments of 128 tokens with a 50% overlap. This procedure enabled the analysis of local contextual patterns across the full lyric while reducing the loss of contextual information caused by sequence length limitations.

Because each lyric was divided into multiple overlapping segments, the outputs generated by the RoBERTuito-based models were aggregated through statistical descriptors to obtain a fixed-length representation for each song. Specifically, the mean, standard deviation, maximum, minimum, median, and 90th percentile were computed across blocks, resulting in 432 additional features per lyric derived from the statistical aggregation of the outputs produced by the sentiment, emotion, hate-speech, and irony models.

Finally, the Qwen contextual embeddings and the RoBERTuito-derived statistical features were concatenated into a unified feature vector. This combined representation served as input to the final ensemble classifiers used in Task 2 and Task 3.

3.5. Feature Selection and Hyperparameter Optimization with Optuna

Optuna [15] was used during the experimental process to optimize model hyperparameters and improve the final decision process. The optimization objective was based mainly on macro F1-score, since this metric is more appropriate for imbalanced classification problems.

Hyperparameter optimization was performed using Optuna for all TF-IDF-based experiments. For Task 2, the optimization process was conducted independently for the binary classifiers associated with each misogyny category, whereas for Task 3 it was applied to the binary gender stereotype classifier.

For the embedding-based experiments, Optuna was additionally employed to perform feature selection. Rather than selecting individual variables, the optimization process operated at the feature-group level. The evaluated groups included Qwen embeddings and RoBERTuito models (sentiment, emotion, hate-speech, and irony), as well as groups defined by different statistical aggregation descriptors (mean, standard deviation, maximum, minimum, median, and 90th percentile). Each group could be independently enabled or disabled during optimization.

Because the final representation combined high-dimensional Qwen embeddings with RoBERTuito-derived linguistic features, feature selection was employed to reduce redundancy and improve model performance. The RoBERTuito-derived features capture different linguistic phenomena, including sentiment, emotion, hate speech, and irony, whose predictive value may vary across tasks and labels. Therefore, Optuna was used to evaluate feature groups jointly with the model hyperparameters, enabling an empirical assessment of their contribution and retaining only those groups that improved validation performance.

3.6. Ensemble-Based Classification

The final methodology used ensemble learning for Task 2 and Task 3. Each model in the ensemble was trained independently using the selected feature representation. During inference, the models produced probability estimates, which were averaged to obtain a combined prediction.

This probability-averaging strategy was selected because it allowed the system to combine complementary decision patterns from different classifiers. Instead of depending on a single model, the ensemble integrated several models that could capture different aspects of the data. This is useful for the MiSonGyny tasks because misogynistic language and gender stereotypes may appear in explicit, implicit, emotional, or contextual forms.

For each ensemble, the final probability was computed as follows:

Table 5
Final ensemble configuration for Task 2 and Task 3.

Task	Label	Models in the ensemble
Task 2	Sexualization	Logistic Regression, SVM with RBF kernel, XGBoost
Task 2	Violence	Logistic Regression, SVM with RBF kernel, LightGBM
Task 2	Hate	CatBoost, LightGBM, SVM with RBF kernel, XGBoost
Task 3	Gender stereotype	Logistic Regression, SVM with RBF kernel, LightGBM

$$P_{final}(y = 1|x) = \frac{1}{K} \sum_{k=1}^{K} P_k(y = 1|x)$$

where K is the number of models in the ensemble, $P_k(y = 1|x)$ is the probability predicted by the k-th model, and $P_{final}(y = 1|x)$ is the averaged probability used to produce the final decision.

For Task 2, one independent ensemble was trained for each label: sexualization, violence, and hate. For Task 3, a separate ensemble was trained for gender stereotype detection. The final ensemble configuration for Task 2 and Task 3 is summarized in Table 5. As shown in the table, each ensemble combines complementary classifiers, including Logistic Regression, SVM with RBF kernel, XGBoost, LightGBM, and CatBoost depending on the target label.

3.7. Threshold Optimization

After obtaining probability estimates from the individual models and ensembles, threshold tuning was applied to optimize the final binary decisions.

Instead of applying a fixed threshold of 0.5 to all labels, the threshold for each binary decision was selected according to the validation performance. The objective was to maximize macro F1-score, allowing the system to adjust the decision boundary according to the behavior of each label. This was useful for labels with class imbalance.

For Task 2, threshold optimization was performed independently for sexualization, violence, and hate. For Task 3, threshold tuning was applied to the averaged probability of the stereotype ensemble. The selected thresholds were then used during final inference on the test data. The final threshold values adopted by the proposed system are reported in Table 6 to facilitate reproducibility.

Table 6
Optimized decision thresholds selected on the validation set.

Task - Label	Threshold
Task 2 - Sexualization	0.554
Task 2 - Violence	0.434
Task 2 - Hate	0.576
Task 3 - Stereotype	0.596

3.8. Final Hierarchical Inference

An overview of the final methodology was presented in Figure 2. The proposed approach followed a hierarchical workflow in which the Longformer classifier first determined whether a lyric is misogynistic. If the lyric is classified as non-misogynistic, the system directly assigns negative labels for Task 2 and Task 3.

For lyrics identified as misogynistic, Qwen embeddings and RoBERTuito-derived linguistic features were extracted and were used as input to the ensemble classifiers. The Task 2 ensembles independently predicted the presence of sexualization, violence, and hate, while the Task 3 ensemble predicts the

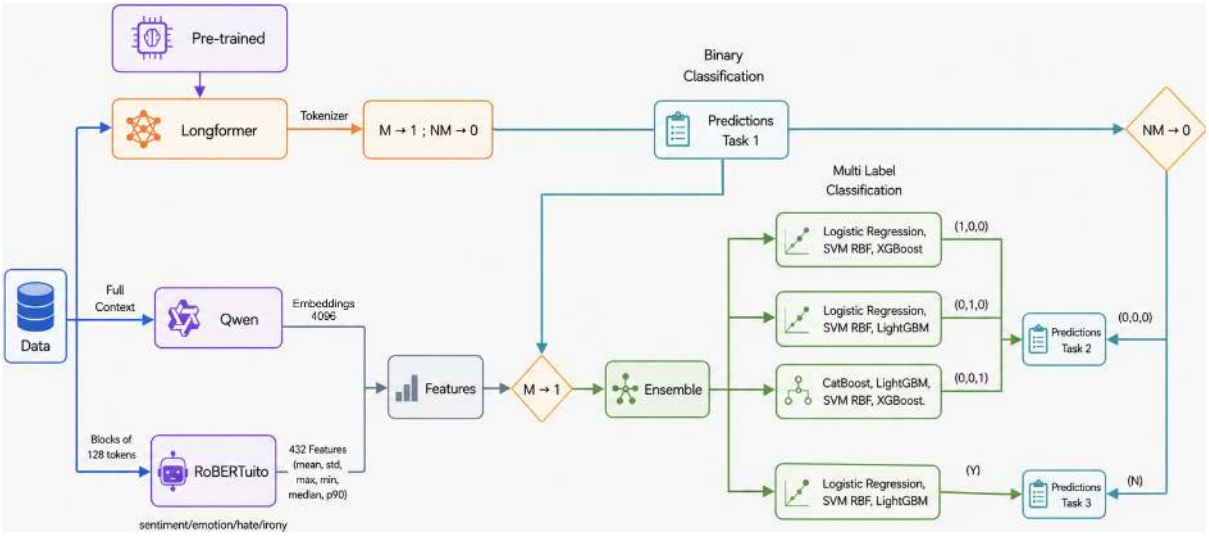

Figure 2: Final ensemble-based methodology used for the MiSonGyny 2026 tasks. Task 1 is addressed with a Longformer-based binary classifier. In parallel, Qwen embeddings and RoBERTuito-based statistical features are extracted and concatenated into a unified feature representation. If Task 1 predicts a non-misogynistic lyric, Task 2 and Task 3 receive negative outputs directly. Otherwise, ensemble classifiers are activated for multilabel misogyny type classification and gender stereotype identification.

presence of gender stereotypes. Final labels were obtained by applying the optimized decision thresholds to the ensemble probabilities.

4. Experiments and Results

This section presents the experimental evaluation of the proposed methodology for the MiSonGyny 2026 shared task. First, we describe the preliminary experiments. Next, we report the official results obtained in the three subtasks and compare them with the internal evaluation. Finally, we discuss the main findings and limitations of the proposed approach.

4.1. Internal Evaluation

During the experimental stage, several machine learning baselines based on TF-IDF representations and different preprocessing configurations were evaluated. For Task 1, the XGBoost model trained with embeddings and RoBERTuito-derived features achieved an F1-score of 0.834, showing competitive performance among the evaluated models. However, the final selected model was a Longformer classifier, which achieved an internal F1-score of **0.845** on the held-out test split and was better suited to process long song lyrics while preserving broader contextual information. Therefore, Longformer was selected as the first decision layer of the hierarchical system.

For Task 2, misogyny type identification was formulated as a multilabel classification problem using a binary relevance strategy. As an initial baseline, three independent RidgeClassifier models were trained to predict sexualization, violence, and hate, achieving an F1-score of 0.639. Additional experiments incorporated Qwen embeddings and RoBERTuito-derived statistical features. Using the final ensemble-based methodology, the internal evaluation conducted on the misogynistic data achieved an F1-score of **0.775**, outperforming the TF-IDF baseline. The reported score corresponds to the macro average across the three misogyny labels.

For Task 3, gender stereotype identification was initially addressed using a DecisionTreeClassifier trained on TF-IDF features, obtaining an F1-score of 0.765. Additional experiments with Qwen embeddings and RoBERTuito-derived statistical features achieved an F1-score of **0.780** in the internal

Table 7
Ablation study of the proposed architecture. Flat uses independent classifiers, Hierarchical applies the proposed two-stage architecture, and Final corresponds to the complete system with ensemble learning, hyperparameter tuning, and threshold tuning.

Task - Label	Setting	Model	Macro F1	Δ
T2 - Sexualization	Flat	LR	0.828	–
	Hierarchical	LR	0.843	+0.015
	Final	Ensemble	**0.849**	+0.021
T2 - Violence	Flat	LR	0.660	–
	Hierarchical	LR	0.680	+0.020
	Final	Ensemble	**0.698**	+0.038
T2 - Hate	Flat	XGB	0.767	–
	Hierarchical	XGB	0.777	+0.010
	Final	Ensemble	**0.779**	+0.012
T3 - Stereotype	Flat	LR	0.757	–
	Hierarchical	LR	0.767	+0.010
	Final	Ensemble	**0.780**	+0.023

evaluation.

Table 7 evaluates the contribution of the proposed hierarchical architecture. To isolate the effect of the hierarchy, the same classifier was used in both the flat and hierarchical configurations for each task. The results show consistent improvements across all downstream tasks when the hierarchical decision process is applied. The largest gain was observed for violence detection (+0.020). Improvements were also obtained for sexualization (+0.015), hate (+0.010), and gender stereotype detection (+0.010). These results indicate that conditioning the downstream classifiers on the misogyny detection stage helps reduce noise from non-misogynistic lyrics and provides modest but consistent improvements. Additional gains were achieved through ensemble learning, hyperparameter tuning and threshold tuning, producing the final system used in the shared task.

To quantify error propagation in the hierarchical framework, we analyzed the false negatives produced by Task 1, since these instances are prevented from reaching the downstream classifiers. The results show that 26.7% of the misogynistic lyrics were incorrectly classified as non-misogynistic and therefore blocked by the first stage of the pipeline. Considering the downstream tasks, this resulted in the loss of 11.2% of the positive sexualization instances, 5.7% of the positive violence instances, 10.4% of the positive hate instances, and 11.8% of the positive Gender Stereotype instances. These findings confirm the presence of error propagation in the hierarchical architecture. However, the impact on downstream labels was substantially lower than the raw false-negative rate observed in Task 1. Furthermore, despite this limitation, the hierarchical formulation consistently outperformed the flat baseline across all downstream tasks (Table 7), indicating that the benefits of filtering non-misogynistic lyrics outweigh the performance degradation caused by the blocked instances.

The confusion matrices in Figures 3, 4, and 5 provide a more detailed view of the internal classification performance. The Task 1 matrix corresponds to the Longformer model evaluated on a hold-out partition, resulting in fewer instances than the other tasks. The matrices for Task 2 and Task 3 were computed only on lyrics labeled as misogynistic, since both classifiers were trained on that subset. Task 1 shows strong performance for both classes. In Task 2, sexualization achieved the best results, while violence remained the most challenging category. Task 3 highlights the difficulty of identifying gender stereotypes, which are often expressed through subtle linguistic patterns.

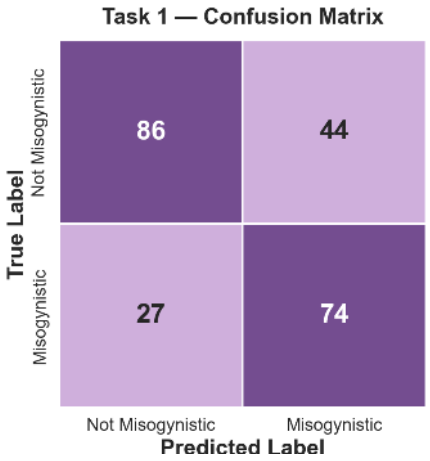

Figure 3: Confusion matrix for Task 1: Misogyny Detection.

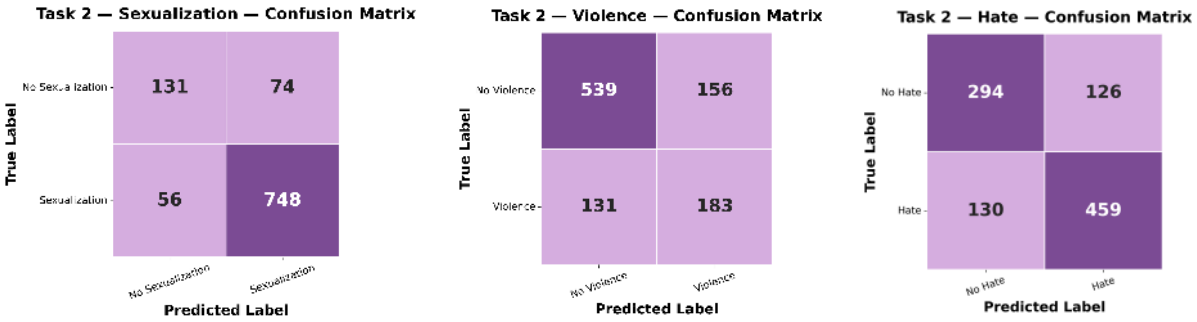

Figure 4: Confusion matrices for Task 2: Sexualization, Violence, and Hate.

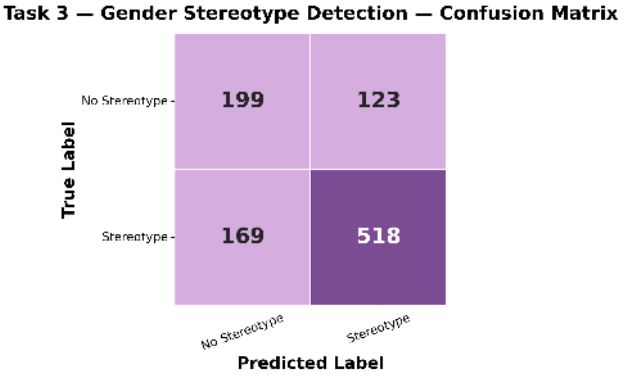

Figure 5: Confusion matrix for Task 3: Gender Stereotype Detection.

4.2. Official Leaderboard Results

The official evaluation of the MiSonGyny 2026 shared task ranked participants according to their best F1-score for each subtask. Tables 8, 9, and 10 show the final leaderboard results for Task 1, Task 2, and Task 3, respectively. In each table, the row corresponding to the cicese-lcdaa team is highlighted in bold.

The final system achieved an F1-score of 0.8533 in Task 1, 0.6846 in Task 2, and 0.7924 in Task 3. These results correspond to the official evaluation of the shared task and reflect the performance of the final consolidated methodology.

Table 8
Official results for Task 1: Misogyny Detection.

Contestant	Submission ID	F1-score	Place
tbernal	706465	0.8856	1
ikanate	712158	0.8712	2
aerojash	713406	0.8581	3
manuelsanz	706362	0.8559	4
vramos	706470	0.8559	4
cicese-lcdaa	**714148**	**0.8533**	**5**
franrodrigo	706467	0.8448	6
arjun108	–	0.8434	7
VerbaNexAI Lab	706460	0.8417	8
sergiomarinnx	713859	0.8334	9
amisadai_ed	714215	0.8219	10
misongyny	706474	0.7931	11

Table 9
Official results for Task 2: Misogyny Type Identification.

Contestant	Submission ID	F1-score	Place
tbernal	706465	0.7385	1
manuelsanz	706362	0.7079	2
vramos	706470	0.6939	3
cicese-lcdaa	**714415**	**0.6846**	**4**
ikanate	707773	0.6813	5
arjun108	–	0.6737	6
aerojash	714495	0.6699	7
amisadai_ed	714215	0.6683	8
franrodrigo	706467	0.6645	9
sergiomarinnx	713859	0.6562	10
misongyny	706474	0.5015	11
VerbaNexAI Lab	706460	0.4022	12

Table 10
Official results for Task 3: Gender Stereotype Identification.

Contestant	Submission ID	F1-score	Place
tbernal	706465	0.8283	1
ikanate	707773	0.8109	2
aerojash	713406	0.8071	3
franrodrigo	706467	0.8039	4
manuelsanz	706362	0.8015	5
vramos	706470	0.8015	5
cicese-lcdaa	**714415**	**0.7924**	**6**
amisadai_ed	714215	0.7860	7
VerbaNexAI Lab	706460	0.7798	8
sergiomarinnx	713859	0.7724	9
misongyny	706474	0.7206	10
arjun108	–	0.6366	11

4.3. Discussion

The proposed methodology achieved competitive performance across the three subtasks of the MiSonGyny 2026 shared task. The strongest relative result was obtained in Task 2, where the system ranked 4th place. This result suggests that the combination of contextual embeddings and linguistic features was particularly effective for multilabel misogyny type identification, a task in which multiple non-exclusive categories may coexist within the same lyric.

For Task 1, the Longformer-based classifier demonstrated the usefulness of long-context transformer architectures for processing song lyrics. The ability to model long textual dependencies appears especially relevant in this domain, where lyrics frequently contain repetitions, figurative language, and contextual references distributed across the song.

Task 3 remained comparatively challenging. Gender stereotypes are often expressed through subtle, implicit, or culturally dependent language patterns rather than explicit offensive expressions, making them more difficult to identify automatically.

A limitation of the proposed framework is its hierarchical structure. Since Task 2 and Task 3 predictions depend on the output of Task 1, errors in the first stage may propagate to the downstream classifiers. In particular, false negatives in Task 1 prevent the system from identifying misogyny types or gender stereotypes in subsequent stages. Despite this limitation, the hierarchical design reduced the number of fine-grained predictions assigned to lyrics classified as non-misogynistic.

5. Conclusions

This paper presented the participation of the CICESE-LCDAA team in the MiSonGyny 2026 shared task, which focuses on the automatic detection and characterization of misogynistic content in Spanish song lyrics. The proposed approach addressed the three subtasks of the challenge through a hierarchical framework that first determines whether a lyric is misogynistic and subsequently predicts misogyny categories and gender stereotypes. This design ensured consistency across the subtasks by restricting fine-grained predictions to lyrics identified as misogynistic.

The experimental results suggest that integrating long-context transformer models with contextual embeddings, linguistic features, and ensemble learning can be beneficial for the automatic analysis of misogynistic content in song lyrics. The proposed system achieved an F1-score of 0.8533 in Task 1 (5th place), 0.6846 in Task 2 (4th place), and 0.7924 in Task 3 (6th place). Notably, the strongest relative performance was obtained in Task 2, the multilabel misogyny type identification task. This result suggests that the combination of Qwen embeddings and RoBERTuito-derived linguistic features provides complementary information that helps identify the coexistence of multiple misogynistic categories within the same lyric.

The results also highlight the benefits of integrating semantic representations with affective and linguistic signals. While contextual embeddings capture the overall meaning of the lyrics, RoBERTuito features contribute complementary information related to sentiment, emotion, hate-related content, and irony. Together, these representations improve the characterization of complex and nuanced forms of misogynistic language.

Despite these promising results, the task remains challenging due to the implicit, figurative, and culturally dependent nature of song lyrics. Additional difficulties arise from the overlap between misogyny categories, the presence of subtle stereotypes, and the multilabel structure of the problem, where multiple forms of misogynistic content may coexist within the same lyric.

Future work will focus on conducting a detailed error analysis to better understand category-specific failure modes, exploring architectures tailored to individual labels, and investigating data augmentation strategies and larger pretrained language models. More broadly, the findings of this work support the use of hybrid approaches that combine large language model embeddings with task-specific linguistic features for the automatic analysis of misogynistic content and gender stereotypes in Spanish-language music.

Acknowledgments

The authors would like to thank the Laboratorio de Ciencia de Datos y Aprendizaje Automático (LCDAA) for the academic, technical, and computational support provided during the development of this work. Special thanks are extended to Irvin Hussein López-Nava for the opportunity to participate in this research project and for his valuable supervision. The author also thanks Joan Manuel Raygoza-Romero for his mentorship, guidance, and continuous support throughout the project. Additionally, the author acknowledges Jesus Antonio Navarrete-Lopez and Cielo Aholiva Higuera-Gutiérrez for their collaboration, feedback, and support during the development of this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: check grammar, spelling and reword. 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] A. Bonet-Jover, J. A. 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.

[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] L. Betti, C. Abrate, A. Kaltenbrunner, Large scale analysis of gender bias and sexism in song lyrics, EPJ Data Science 12 (2023) 10. URL: https://doi.org/10.1140/epjds/s13688-023-00384-8. doi:10.1140/epjds/s13688-023-00384-8.

[4] United Nations, What is hate speech?, 2026. URL: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech, accessed: 2026-05-09.

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

[6] 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).

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[10] 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, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] 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, arXiv preprint arXiv:1907.11692 (2019).

[12] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[13] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025).

[14] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: 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. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785/.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.