CorpusDelicti-UC3M at MiSonGyny 2026: Multi-Model Voting Ensembles with Chunk Aggregation for Misogynistic Lyric Detection

Victor Ramos-Osuna[1,*], Manuel Sanz-Muñoz[2], Emanuel Pacheco-Rojas[2] and Yulicenia Díaz-Cabrera[2]

[1]Universidad Politécnica de Madrid, Department of Computer Systems, C. de Alan Turing, s/n, 28031, Madrid, Spain
[2]Universidad Carlos III de Madrid, Madrid, Spain

Abstract

Music plays a powerful role in popular culture, yet it can also normalize misogynistic attitudes and harmful gender stereotypes through derogatory or objectifying lyrics. This paper presents the participation of the CorpusDelicti-UC3M team in the MiSonGyny 2026 shared task, which addresses the automatic detection of misogyny and stereotypes in Spanish song lyrics. To address Transformer input-length limitations and the long, unstructured nature of song lyrics, the system uses overlapping sliding windows, sampling a random window from each song during training to improve generalization and reduce overfitting to specific lyrical segments, while classification is performed through ensemble architectures that combine Spanish-specific and multilingual models. Segment predictions are aggregated using the maximum probability across windows, with validation-based threshold tuning to maximize macro F1-score. The approach further examines the effect of task-specific formatting, specialized loss functions, and targeted data enrichment. Subtasks 1 and 2 benefited from punctuation-based lyric structuring, while the multilabel Subtask 2 further employed a One-vs-Rest (OvR) hard-voting framework with hybrid Gemma-based paraphrasing and AEDA augmentation to address severe class imbalance.

Keywords

Misogyny Detection, Spanish Lyrics, Transformer Ensembles, Sliding Window Aggregation, One-vs-Rest Classification, Data Augmentation, Natural Language Processing

1. Introduction

Music is a pervasive cultural force that significantly shapes public perceptions and social dynamics. While it serves as a powerful medium for artistic expression, it can also reflect and amplify harmful societal discourses. Song lyrics frequently normalize misogynistic attitudes, sexual objectification, and gender-based violence by embedding derogatory stereotypes within rhythmic and poetic structures [1]. Identifying and addressing these harmful narratives is essential to mitigating the perpetuation of gender inequality, making the application of Natural Language Processing (NLP) to detect hate speech and sexism in creative domains an increasingly active area of research.

Building upon previous efforts to identify harmful content in Iberian languages [2], the MiSonGyny 2026 shared task [3, 4] challenges participants to automatically detect and classify misogynistic speech in Spanish song lyrics. This task is especially relevant given the global cultural reach of Spanish-language music and extends prior research on detecting harmful content targeting specific groups—such as the LGBT+phobic content addressed in the HOMO-MEX task [5]—into the nuanced domain of fine-grained misogyny.

Processing song lyrics presents unique technical challenges for modern NLP architectures. Lyrical texts are inherently unstructured, frequently employing slang, metaphors, and figurative language that standard text classifiers struggle to interpret. Furthermore, song lyrics often exceed the 512-token input

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
victor.ramos.osuna@upm.es (V. Ramos-Osuna); 100472082@alumnos.uc3m.es (M. Sanz-Muñoz); 100575869@alumnos.uc3m.es (E. Pacheco-Rojas); 100570982@alumnos.uc3m.es (Y. Díaz-Cabrera)
0009-0009-6139-967X (V. Ramos-Osuna)

limit of standard Transformer models such as BETO [6], and naive truncation risks discarding crucial context where harmful language may be dispersed across the full text.

This paper describes the methodology of the CorpusDelicti-UC3M team. Our system addresses these challenges through a pipeline centered on an overlapping sliding window strategy for inference combined with randomized window sampling during training, max-probability chunk aggregation, dynamic validation threshold tuning, and multi-model voting ensembles over a diverse set of Spanish-specific and multilingual Transformer architectures. The multilabel nature of Subtask 2 is specifically addressed via a One-vs-Rest (OvR) framework supported by a hybrid data augmentation strategy.

2. Task Description

The MiSonGyny 2026 shared task [4] is structured into three subtasks. **Subtask 1** addresses binary misogyny detection at the song level, classifying lyrics as either Misogynistic (M) or Non-Misogynistic (NM). **Subtask 2** requires fine-grained misogyny type classification as a multilabel problem, predicting the simultaneous presence of Sexualization (S), Violence (V), and/or Hate (H). **Subtask 3** tackles stereotype identification, distinguishing lyrics that depict or reinforce generalized gender-role beliefs (Stereotypical) from those that do not (Non-Stereotypical).

3. Related Work

3.1. Sexism and Misogyny Detection in Iberian Languages

The automatic detection of hate speech and sexist language has become a critical research area within NLP [7]. Early efforts in Spanish-language contexts were spearheaded by the EXIST shared tasks, which established benchmarks for identifying both explicit and subtle sexist behaviors in social media and emphasized the subjective complexity of sexism as a linguistic phenomenon [2]. The HOMO-MEX task at IberLEF 2024 further extended this line of work to LGBT+phobia detection in Mexican Spanish, including a song lyrics subtask [5]. The first MiSonGyny edition in 2025 provided the first dedicated corpus for misogynistic speech in Spanish song lyrics, defining binary detection and fine-grained multilabel categories [1].

3.2. Prior Approaches to MiSonGyny

Two participating teams from the 2025 edition are particularly relevant to our work, as both share key architectural choices with our system. Toledo and Segura-Bedmar [8] combined Transformer ensembles with instruction-tuned generative models, exploring how large language model outputs can complement discriminative classifiers for misogyny detection. Díez-Fenoy et al. [9] employed a Transformer ensemble approach based on BERT-family models for Spanish, demonstrating the effectiveness of model combination over individual fine-tuned classifiers.

3.3. Transformer Architectures for Spanish NLP

The performance of NLP systems has been significantly enhanced by Transformer models pre-trained on large-scale language-specific corpora [10, 11]. In the Spanish domain, BETO demonstrated that models trained exclusively on Spanish text consistently outperform multilingual baselines across NLU tasks [6]. RoBERTuito extended this to informal social media text, sharing linguistic properties—slang and colloquialisms—with song lyrics [12]. For extended-context tasks, the Spanish Longformer processes sequences well beyond the standard 512-token limit [13], while XLM-R provides strong cross-lingual representations [14].

3.4. Long Text Processing and Chunking Strategies

A core bottleneck in standard Transformer architectures is their fixed maximum input length, frequently exceeded by full-length song lyrics. Naive truncation risks discarding discriminative information distributed across the document [15]. Sliding window and multi-passage normalization techniques process texts in overlapping segments to preserve contextual coherence [16]. Random window sampling during training acts as an implicit form of data augmentation, ensuring the model observes diverse portions of each text across training steps rather than memorizing position-specific cues. At inference time, the BERT-MaxP aggregation strategy [16] motivates max-probability aggregation: relevant signal may appear at any position in the document, and the maximum-confidence chunk score best captures it.

3.5. Ensemble Methods and Data Augmentation

Ensemble learning mitigates the high variance of individual fine-tuned models, particularly on small specialized datasets [11]. Hard voting combines discrete class predictions by majority rule, while soft voting averages class probability distributions—providing finer-grained aggregation when models produce well-calibrated outputs [17]. For multilabel classification, the One-vs-Rest (OvR) framework decomposes the problem into independent binary classifiers per label and remains a competitive strategy, particularly when label correlations are weak [18]. Data augmentation strategies address class imbalance and data scarcity: AEDA randomly inserts punctuation to shift token positions without altering semantics [19], while LLM-based paraphrasing generates semantically equivalent synthetic examples that preserve the original label [20, 21].

3.6. Positioning

While prior work established strong individual baselines, several open challenges remain unaddressed in the MiSonGyny setting. Existing systems did not explicitly handle lyrical sequences exceeding standard Transformer context windows, nor apply dynamic threshold optimization to combat class imbalance. The systematic combination of diverse Spanish-specific and multilingual architectures within an ensemble framework tailored to the lyrical domain remains underexplored, as does per-subtask augmentation targeting the severe minority-label skew present in fine-grained misogyny classification. Our work directly addresses these gaps through a unified pipeline integrating all of the above strategies.

4. Data

4.1. Dataset Description

The MiSonGyny 2026 corpus comprises a diverse collection of Spanish song lyrics annotated for misogynistic and stereotypical language. We utilized the provided training split of 2,303 unique songs, mapping any missing labels to their respective negative classes. Table 1 summarizes the class distributions across subtasks. Subtask 2 exhibits the most severe imbalance, with Violence appearing in only 13.6% of songs, which directly motivates the augmentation strategy described in Section 5.6. The dataset was partitioned into a 70/15/15 ratio for training, validation, and evaluation.

4.2. Cross-Edition Dataset Analysis and Label Shift

To further contextualize the task and explore potential data augmentation strategies, we performed a historical collision analysis against the previous edition's dataset [1]. To account for transcription inconsistencies across editions (e.g., varying punctuation, diacritics, or missing verses), we applied a robust heuristic matching pipeline. Textual inputs from both datasets were lowercased, unicode-normalized to remove diacritics, stripped of punctuation, and tokenized. A Jaccard similarity metric was then computed over the resulting token sets.

Table 1
Class distributions across the three subtasks within the 2,303-song training split. Missing labels are mapped to the negative class.

Subtask	Label Category	Positive Instances	Negative Instances
Subtask 1	Misogynistic (M vs. NM)	1,009 (43.8%)	1,294 (56.2%)
Subtask 2	Sexualization	804 (34.9%)	1,499 (65.1%)
	Violence	314 (13.6%)	1,989 (86.4%)
	Hate	589 (25.6%)	1,714 (74.4%)
Subtask 3	Gender Stereotype	687 (29.8%)	1,616 (70.2%)

Using a similarity threshold of ≥ 0.60, selected to balance robustness to transcription noise with the avoidance of spurious matches, we identified 1,244 overlapping lyrics between the legacy dataset (2,631 total instances) and the 2026 training split (2,303 instances). Notably, our analysis revealed a significant label shift, as 207 of the matched songs (approximately 16.6% of the overlapping subset) underwent a binary label inversion (Misogynistic to Non-Misogynistic, or vice versa) in the 2026 release. This substantial shift underscores the subjective, metaphorical nature of musical lyrics and suggests a rigorous refinement of the annotation guidelines by the task organizers to correct legacy noise or redefine borderline instances.

Critically, this label shift raised concerns extending beyond the directly matched instances. Given that annotation criteria had demonstrably evolved between editions, we could not confidently assume that unmatched legacy samples (those absent from the 2026 training split) conformed to the current decision boundaries. A non-duplicated legacy sample previously labeled as Misogynistic, for instance, might well have been reclassified as Non-Misogynistic under the 2026 guidelines, had it appeared in the new release. To mitigate the risk of training our models on potentially deprecated classification logic, we therefore opted to exclude the full legacy dataset from our augmentation pipeline, forgoing the additional signal in favour of label integrity.

5. Methodology

5.1. Long Sequence Handling

Song lyrics routinely exceed the maximum input lengths of standard Transformer models. Rather than truncating sequences, we apply a unified strategy across all models and subtasks.

Training: Randomized Window Sampling. During training, a data collator extracts a random contiguous crop of up to `MAX_LEN` tokens from each lyric at every batch. This means that across training epochs the model is exposed to different portions of each song, acting as an implicit data augmentation that prevents overfitting to specific lyrical segments and improves generalization to the full lyrical context.

Inference: Overlapping Sliding Window with Max Aggregation. At inference time, each lyric is parsed exhaustively using a sliding window of length `MAX_LEN` tokens and a stride of `MAX_LEN` // 2, producing overlapping chunks that preserve near-boundary context. Each chunk is independently classified, yielding a probability or logit score per class. The final document-level prediction is obtained by max-pooling across all chunks—selecting the highest triggered probability for each class across all windows. This follows the BERT-MaxP aggregation rationale [16]: the most informative signal for a label may appear anywhere within the lyric, and the maximum-confidence chunk best captures it.

5.2. ML Baselines

To establish a classical baseline and contextualize downstream deep-learning results, we evaluate three classical classification algorithms: Logistic Regression, Random Forest, and Support Vector Machines (SVM). For the binary classification formulations in Subtask 1 (misogyny detection) and Subtask 3 (stereotype detection), the base estimators are applied directly. Conversely, to handle the multilabel configuration of Subtask 2 (simultaneous prediction of Sexualization, Violence, and Hate), we deploy a native scikit-learn [22] One-vs-Rest (`OneVsRestClassifier`) framework. This wrapper fits an independent meta-estimator per target label category.

Text representations across all classical pipelines are generated uniformly. Lyrical inputs are lowercased, tokenized, filtered to remove Spanish stop words and punctuation elements, and lemmatized utilizing the `es_core_news_sm` pipeline via spaCy. Tokens spanning fewer than three characters are pruned. The preprocessed text structures are subsequently transformed into numeric vectors via a Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer restricted to the top 5,000 maximum features.

Hyperparameters for each classifier are optimized using a genetic algorithm-based space exploration framework provided by the `sklearn-genetic-opt` library via the `GASearchCV` module [23]. Optimization runs across a population size of 20 over 10 evolving generations, evaluating candidate configurations with a 3-fold stratified cross-validation setup explicitly maximizing macro F_1-score. The specific hyperparameter search boundaries established per algorithm include:

- **Logistic Regression:** A continuous optimization space for the inverse regularization strength parameter $C \in [10^{-3}, 10^{2}]$ drawn from a log-uniform distribution, evaluated across both balanced and unweighted (`None`) class weight assignments.
- **Random Forest:** Intersecting integer ranges restricting the number of trees (`n_estimators` $\in [50, 500]$), maximum tree depth (`max_depth` $\in [5, 50]$), and the minimum samples required to trigger an internal node split (`min_samples_split` $\in [2, 11]$). Class distributions are penalized using a fixed `balanced` configuration.
- **Support Vector Machine:** Continuous exploration of the margin regularization parameter $C \in [10^{-3}, 10^{2}]$ log-uniformly distributed, cross-evaluated against both Linear and Radial Basis Function (`rbf`) kernel configurations with class balancing turned on.

Both the original baseline un-preprocessed textual fields and the structurally optimized punctuation-mapped variants are evaluated across all subtasks to provide a comprehensive baseline.

5.3. Transformer Benchmarking

We benchmark six Transformer architectures across all subtasks under a uniform training configuration, varying only the architecture-specific context length and learning rate (Table 2):

- **BETO** [6] (`dccuchile/bert-base-spanish-wwm-cased`): A Spanish-only Transformer model based on the BERT architecture, pre-trained on a massive corpus comprising the Spanish portion of Wikipedia and the OPUS project. We utilized `max_len`=512 and a learning rate of 5×10^{-6}.
- **DistilBETO** [24] (`dccuchile/distilbert-base-spanish-uncased`): A distilled version of Spanish BERT available via HuggingFace. Knowledge distillation reduces the parameter count and computational overhead while aiming to preserve the performance of the original teacher model. Training parameters were set to `max_len`=512 and lr=2×10^{-5}.
- **RoBERTuito** [12] (`pysentimiento/robertuito-base-uncased`): A RoBERTa-based language model specifically crafted for user-generated text, pre-trained on over 500 million Spanish tweets. We configured it with `max_len`=128 and lr=2×10^{-5}. Despite its shorter context window, its specialized pre-training on informal, unstructured social media text enables it to perform competitively when integrated with our sliding window strategy.

- **MarIA** [10] (`IsGarrido/roberta-base-bne`): A RoBERTa-based model pre-trained via Masked Language Modeling on a massive 570 GB corpus of clean, deduplicated text from the Spanish Web Archive. This provides robust, general-domain Spanish representations. We used `max_len`=512 and lr=2×10^{-5}.
- **XLM-R** [14] (`xlm-roberta-base`): A highly scalable multilingual RoBERTa model trained on more than two terabytes of filtered CommonCrawl data across 100 languages. Its inclusion brings robust cross-lingual representations to the ensemble. We applied `max_len`=512 and lr=2×10^{-5}.
- **Longformer** [13] (`mrm8488/longformer-base-4096-spanish`): A Spanish-adapted extended-context model designed to bypass the standard 512-token barrier typical of traditional Transformers. This architecture is particularly well-suited for processing the entirety of long song lyrics. We fine-tuned it with `max_len`=1024 and lr=2×10^{-5}.

All models are trained with AdamW (fused), a linear learning rate scheduler, a maximum of 10 epochs with early stopping (patience=3), batch size 16 with gradient accumulation steps of 2 (reduced to batch size 8 with 4 accumulation steps for memory-intensive configurations), dropout of 0.1 across the classifier head, attention, and hidden layers, weight decay of 0.01, and warmup ratio of 0.1. The data is split 70/15/15 into train, validation, and evaluation sets. For Subtasks 1 and 2, standard cross-entropy loss is optimized. For Subtask 3, class weights computed from the original label distributions are applied via a weighted cross-entropy loss function during training to explicitly mitigate severe class imbalance. The held-out evaluation set provides the reported benchmark scores.

Table 2
Model-specific hyperparameter configurations.

Model	Max Len	Learning Rate	Batch / Accum
BETO	512	5×10^{-6}	16 / 2
DistilBETO	512	2×10^{-5}	16 / 2
RoBERTuito	128	2×10^{-5}	16 / 2
MarIA	512	2×10^{-5}	16 / 2
XLM-R	512	2×10^{-5}	8 / 4
Longformer	1024	2×10^{-5}	8 / 4

5.4. Validation Threshold Tuning

To combat class imbalance and move beyond the naive 0.5 decision boundary, explicit classification threshold tuning was conducted during the benchmarking phase. For each model, predicted probabilities were extracted from the validation set. A grid search over the range $[0.01, 0.99]$ with a step size of 0.01 was executed to identify the exact classification threshold that maximized the macro F1-score. This optimized model-specific threshold was retained to map final inference probabilities to discrete class assignments. Furthermore, for the soft voting configurations utilized in Subtask 1, the optimal thresholds discovered across all constituent ensemble models were averaged to establish a stable, calibrated ensemble decision boundary.

5.5. Text Preprocessing

Based on empirical evaluation, text preprocessing was applied selectively by subtask rather than uniformly across the full system. Since the main models are BERT-based Transformer architectures, preprocessing was kept lightweight and aligned with the expected input format of pretrained language models. Therefore, aggressive normalization steps such as stop-word removal, lemmatization, stemming, or extensive text cleaning were not applied.

For **Subtasks 1 and 2**, lyric structure was encoded through punctuation mapping: verse boundaries were replaced with commas, while stanza boundaries were replaced with periods. This transformation

was designed to preserve part of the compositional structure of songs within a flat token sequence, without requiring architectural modifications.

The preprocessing pipeline was empirically evaluated with preprocessing enabled and disabled. Results showed that punctuation-based structuring improved performance for **Subtasks 1 and 2**. In contrast, **Subtasks 3** achieved its best results when preprocessing was disabled. Consequently, the baseline unprocessed text was used as-is for stereotype detection.

Additional preprocessing strategies, including contraction expansion and deduplication, were also tested, but they did not improve performance and were therefore excluded from the final configuration.

5.6. Data Augmentation

To address severe class imbalance in Subtask 2, we applied targeted augmentation to the positive instances of each fine-grained misogyny class. The augmented data combines two techniques in an 80/20 split:

- **Generative Paraphrasing (80%):** We used the instruction-tuned `google/gemma-4-E4B-it` model [20] to generate semantically equivalent rewrites of each lyric. A zero-shot prompt enforced strict meaning preservation and suppressed conversational filler:

  > *"Parafrasea la siguiente letra de canción en español naturalmente, manteniendo el mismo significado. Devuelve solo el texto parafraseado, sin explicaciones ni formato adicional, y nada más que el texto parafraseado: [TEXT]"*

- **AEDA (20%):** The remaining samples were generated using the An Easier Data Augmentation technique [19], which randomly inserts punctuation marks into text. This perturbs token positions in the attention layers without changing word order or meaning [25, 21].

5.7. Ensemble Architectures and Submission Configurations

The top-3 performing models per subtask (as determined on the held-out evaluation set) are combined into voting ensembles. Chunk-level max aggregation is applied within each model prior to ensemble combination.

Subtask 1 Configuration (Mean Soft Voting). DistilBETO + BETO + RoBERTuito, trained with standard cross-entropy loss on processed data. Model-level softmax probabilities are averaged across the ensemble, and class assignment is decided by evaluating this mean distribution against the averaged optimized validation threshold. This leverages confidence calibration across well-performing Spanish-domain models.

Subtask 2 Configuration (OvR Hard Voting). DistilBETO + Longformer + RoBERTuito, trained with standard cross-entropy loss on processed data with targeted augmentation. The multilabel Subtask 2 is decomposed into three independent binary classification problems—Sexualization vs. rest, Violence vs. rest, and Hate vs. rest—using the One-vs-Rest framework. The same three base models are fine-tuned separately for each binary label. Predictions are binarized using each model's tuned validation threshold. Hard voting aggregates binary predictions across the three models per label: a label is assigned if the majority of the three models predict it as positive. This is illustrated in Figure 2.

Subtask 3 Configuration (Weighted Hard Voting). RoBERTuito + XLM-R + Longformer, trained with weighted cross-entropy loss on unprocessed baseline data. Class weights penalize errors on underrepresented distributions. Class binarization follows individual tuned validation thresholds, and hard voting determines the final class by majority prediction. This configuration is illustrated in Figure 1.

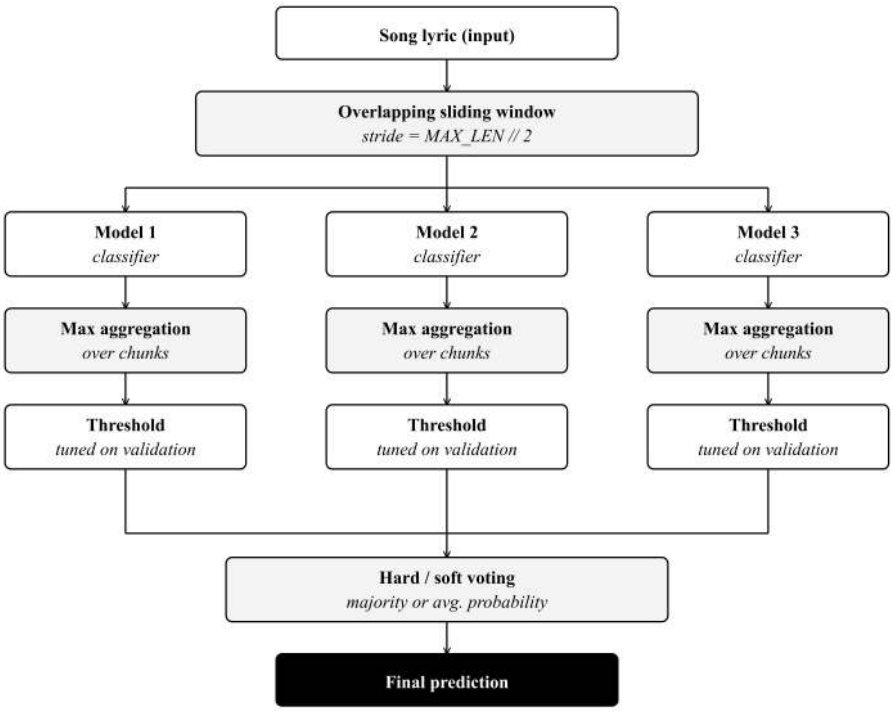

Figure 1: Inference pipelines for Subtasks 1 and 3. Long lyrics are segmented into overlapping windows; each model independently classifies all chunks and aggregates via max pooling; optimized validation thresholds are applied before final voting combinations.

5.8. Final Prediction Training

For final submission predictions, training strategies were adapted to maximize data utilization. Across **all subtasks**, all available data (train, validation, and evaluation splits) were merged into a single training set. To satisfy the HuggingFace `Trainer` API evaluation requirement, a minimal 10-row dummy slice was used as the evaluation set, ensuring no meaningful reduction in the effective training data. For **Subtasks 1 and 3**, the models were trained for the exact number of epochs determined by early stopping during the validation phase. For **Subtask 2**, the number of training epochs was set to the early-stopped epoch count plus one, to better approximate the convergence point on the expanded dataset.

6. Experiments

6.1. ML Baseline Results

Table 3 reports the performance of GA-tuned classical classifiers on the held-out evaluation set across all three subtasks, on both baseline (unprocessed) and preprocessed text variants. Despite the simplicity of the TF-IDF representations, Random Forest achieves competitive macro F1 scores on Subtasks 1 and 3, underscoring the signal present in unigram lexical features. However, all classical methods fall substantially short of the Transformer-based results described below, particularly on Subtask 2, where the multilabel structure and minority label imbalance expose the limitations of bag-of-words representations.

6.2. Transformer Benchmark Results

Tables 4, 5, and 6 report individual model benchmark scores (F1-macro) across subtasks, incorporating the tuned validation thresholds. The top-3 models per subtask are used to define the final ensemble

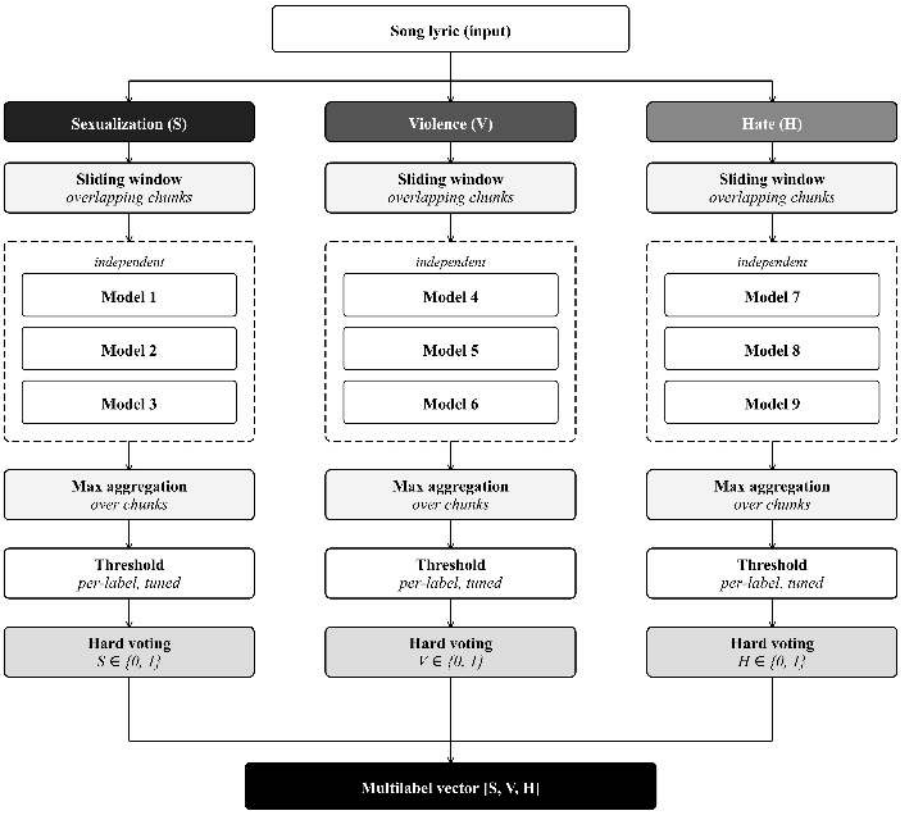

Figure 2: Inference pipeline for Subtask 2. Three independent OvR ensembles (one per label: Sexualization, Violence, Hate) each apply sliding window chunking, per-model max aggregation, threshold binarization, and hard voting to produce a binary decision per label, yielding the final multilabel prediction vector.

Table 3
Traditional ML benchmark results across subtasks (F1-macro). B = Baseline, P = Processed.

	Subtask 1		Subtask 2		Subtask 3	
Model	**B**	**P**	**B**	**P**	**B**	**P**
Random Forest	0.787	0.775	0.551	0.560	0.765	0.793
SVM	0.774	0.770	0.547	0.553	0.721	0.737
Logistic Regression	0.765	0.759	0.575	0.561	0.738	0.741

configuration.

6.3. Ensemble and Final Submission Results

Table 7 reports the F1-macro scores of our primary unified submission on the official shared task test set, alongside the final competition rankings.

7. Discussion

Our results demonstrate that the combination of randomized window sampling during training and max-probability chunk aggregation at inference is an effective strategy for handling long lyrical texts, consistently outperforming the classical ML baselines across all three subtasks. The performance gap between neural models and GA-tuned TF-IDF classifiers confirms that contextual representations capture aspects of misogynistic language in lyrics—such as figurative expression, slang, and structural patterns—that bag-of-words approaches cannot model.

Table 4
Transformer benchmark results for Subtask 1 (F1-macro). B = Baseline, P = Processed, A = Augmented.

Model	B	P	A
BETO	0.854	0.852	0.873
DistilBETO	0.834	0.858	0.828
RoBERTuito	0.802	0.854	0.841
MarIA	0.834	0.825	0.839
XLM-R	0.838	0.835	0.854
Longformer	0.816	0.817	0.820

Table 5
Transformer benchmark results for Subtask 2 (F1-macro per label and overall). B = Baseline, P = Processed, A = Augmented.

Model	B	P	A
BETO	0.584	0.604	0.639
DistilBETO	0.637	0.631	0.625
RoBERTuito	0.609	0.585	0.626
MarIA	0.592	0.602	0.623
XLM-R	0.597	0.584	0.630
Longformer	0.626	0.603	0.626

Table 6
Transformer benchmark results for Subtask 3 (F1-macro). Models trained with weighted cross-entropy. B = Baseline, P = Processed, A = Augmented.

Model	B	P	A
BETO	0.770	0.776	0.790
DistilBETO	0.784	0.778	0.787
RoBERTuito	0.786	0.759	0.788
MarIA	0.761	0.769	0.757
XLM-R	0.785	0.703	0.778
Longformer	0.785	0.767	0.780

Table 7
Final submission results on the MiSonGyny 2026 test set, including final competition rankings.

Component	Strategy	F1-Macro	Ranking
Subtask 1	Soft Voting + Avg. Threshold	0.8559	4th
Subtask 2	OvR Hard Voting + Augmentation	0.7079	2nd
Subtask 3	Weighted Hard Voting (Baseline)	0.8015	5th

Among individual models, RoBERTuito performs competitively despite its restricted 128-token context window. This can be attributed to its pre-training on informal social media Spanish, which shares lexical and stylistic properties with song lyrics, and to the sliding window strategy that compensates for its shorter absolute context. The inclusion of BETO [6] provides robust general-domain Spanish representations as an anchor to the ensemble, proving highly reliable at mitigating the noise introduced by shorter colloquial fragments.

Crucially, validating the threshold space via explicit grid sweeps rather than defaulting to 0.5 proved to be one of the most vital adjustments in balancing class precision and recall under severe skew. Averages derived from these thresholds stabilized the soft voting probabilities used inside Subtask 1's submission workflow.

The success of the OvR framework on Subtask 2—achieving 2nd place—suggests that the three fine-grained misogyny categories (Sexualization, Violence, Hate) are sufficiently independent to benefit from label-specific binary classifiers rather than joint multilabel prediction. The targeted augmentation strategy, applying Gemma 4 paraphrasing to the minority class, proved critical for handling the severe label imbalance in this subtask.

For Subtask 3, the explicit integration of weighted cross-entropy loss provided consistent optimization gains over standard cross-entropy configurations, validating its specific reservation for this subtask. The contrast between Subtasks 1–2 (benefiting from punctuation-based text structuring) and Subtask 3 (performing best on raw text) highlights that the nature of the classification target matters: structural lyric formatting aids in detecting explicit misogynistic content, but may introduce noise when detecting the more subtle implicit stereotypes targeted by Subtask 3.

8. Limitations

Several constraints limited the scope of our experimental exploration. First, due to compute time restrictions, hyperparameter optimization for the Transformer models relied on a fixed uniform configuration across all architectures rather than per-model or per-task tuning. A more principled approach using tools such as Optuna [26] for automated hyperparameter search could yield further improvements, particularly in learning rate and regularization sensitivity across the different model scales. Second, the OvR ensemble for Subtask 2 was constructed using the same top-3 models across all three binary classifiers (Sexualization, Violence, Hate), rather than selecting the optimal model set independently per label. Given the differing difficulty and class balance of each binary subproblem, per-label model selection could improve performance. Third, label correlations in Subtask 2 are not exploited by the OvR decomposition; approaches such as classifier chains could model label dependencies but were outside the scope of the current work due to the same compute constraints.

9. Conclusion

We presented the CorpusDelicti-UC3M system for the MiSonGyny 2026 shared task on misogyny detection in Spanish song lyrics. Our approach addresses the key challenges of long lyrical texts, class imbalance, and multilabel prediction through a unified pipeline of randomized window training, overlapping sliding window inference with max-probability chunk aggregation, and multi-model voting ensembles. The system achieved 2nd place in Subtask 2, 4th place in Subtask 1, and 5th place in Subtask 3. Future work will explore instruction fine-tuning and parameter-efficient adaptation (e.g., LoRA) of large language models for this domain, which may better capture the implicit and figurative nature of misogynistic language in creative texts.

10. Declaration on Generative AI

Generative AI tools were used during the preparation of this paper for syntax and grammar correction, translation assistance, and coding support. All AI-generated suggestions were reviewed and revised by the authors, who take full responsibility for the final content of this work.

References

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

[2] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.
[3] A. Bonet-Jover, J. 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, 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).
[5] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macías, Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.
[6] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[7] A. Vetagiri, P. Pakray, A. Das, A deep dive into automated sexism detection using fine-tuned deep learning and large language models, Engineering Applications of Artificial Intelligence 145 (2025) 110167.
[8] J.-S. Toledo, I. Segura-Bedmar, Hulat_uc3m at misongyny 2025: Detecting misogyny in song lyrics with transformer ensembles and instruction-tuned generative models (2025).
[9] C. Dıez-Fenoy, A. López-González, J. Valle-Dıaz, Dackers at misongyny 2025: A transformer ensemble approach for misogyny detection in spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.
[10] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.
[11] H. Zhang, M. O. Shafiq, Survey of transformers and towards ensemble learning using transformers for natural language processing, Journal of big Data 11 (2024) 25.
[12] 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: 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.
[13] M. Romero, Spanish longformer by manuel romero, https://huggingface.co/mrm8488/longformer-base-4096-spanish, 2022.
[14] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[15] A. Jaiswal, E. Milios, Breaking the token barrier: Chunking and convolution for efficient long text classification with bert, arXiv preprint arXiv:2310.20558 (2023).
[16] Z. Dai, J. Callan, Deeper text understanding for ir with contextual neural language modeling, in: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 985–988.
[17] K. Kyritsis, C. M. Liapis, I. Perikos, M. Paraskevas, V. Kapoulas, From transformers to voting ensembles for interpretable sentiment classification: A comprehensive comparison, Computers 14 (2025) 167.
[18] V. Ashwinkumar, P. P. Arage, R. Jeya, P. Sudhakaran, One-vs-rest vs. voting classifiers for multi-label text classification: An empirical study, in: E3S Web of Conferences, volume 491, EDP Sciences, 2024, p. 01014.

[19] A. Karimi, L. Rossi, A. Prati, Aeda: An easier data augmentation technique for text classification, in: Findings of the association for computational linguistics: EMNLP 2021, 2021, pp. 2748–2754.
[20] Gemma Team, Gemma 4 model card: `google/gemma-4-E4B-it`, https://huggingface.co/google/gemma-4-E4B-it, 2026. Accessed: 2026-05-26.
[21] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, L. A. Tuan, S. Joty, Data augmentation using llms: Data perspectives, learning paradigms and challenges, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 1679–1705.
[22] 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.
[23] R. Arenas, Sklearn-genetic-opt: Hyperparameter tuning and feature selection with genetic algorithms, https://github.com/rodrigo-arenas/Sklearn-genetic-opt, 2026. Accessed: 2026-05-20.
[24] DCC UChile, distilbert-base-spanish-uncased, https://huggingface.co/dccuchile/distilbert-base-spanish-uncased, 2023. Accessed: 2026-05-27.
[25] Y. Chai, H. Xie, J. S. Qin, Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities, Artificial Intelligence Review 59 (2026) 35.
[26] 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 & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.