IReL_IIT(BHU) at MiSonGyny 2026: Consistency Constrained Fine-Tuning for Misogyny Detection in Spanish Song Lyrics

Arjun Mukherjee[1,*,†], Sukomal Pal[1,†]

[1]Indian Institute of Technology, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Abstract

We present the system submitted by IReL_IIT(BHU) to MiSonGyny 2026, a shared task on misogyny detection in Spanish-language song lyrics at IberLEF 2026. Our approach fine-tunes BETO (`dccuchile/bert-base-spanish-wwm-cased`), a BERT-base model pre-trained on a large Spanish corpus, on three subtasks: (1) binary misogyny detection (*M*/*NM*), (2) multilabel misogynistic type classification (Sexualization, Violence, Hate), and (3) binary gender stereotype identification (*Y*/*N*). Task-specific linear classification heads are trained with weighted Binary Cross-Entropy loss to handle class imbalance. We apply a dual-signal early stopping strategy monitoring both validation macro-F1 stagnation and train/validation loss divergence. Cross-task prediction consistency is enforced by post-processing: samples predicted as non-misogynistic in Subtask 1 automatically receive null labels in Subtasks 2 and 3.

Keywords

misogyny detection, hate speech, Spanish NLP, BETO, BERT, multilabel classification, gender stereotypes, song lyrics, IberLEF 2026, MiSonGyny 2026

1. Introduction

Misogyny in music is a pervasive yet underexplored form of gender-based hate speech. Song lyrics, particularly in certain popular genres, frequently contain language that degrades women through explicit sexualisation, threats of violence, and expressions of contempt. Automated detection of such content at scale is critical for content moderation, media analysis, and the broader study of gender-based discrimination in popular culture.

This paper describes the system submitted by IReL_IIT(BHU) to MiSonGyny 2026 [1], the shared task on Misogyny Speech Detection in Spanish Language Song Lyrics, held at IberLEF 2026 [2]. The task poses three classification challenges operating over Castilian (`es_ES`) and Latin American (`es_LATAM`) Spanish lyrics, building on prior work including MiSonGyny 2025 [3] and HOMO-MEX 2024 [4].

Our approach fine-tunes BETO [5], a BERT-based [6] language model pre-trained on a large Spanish corpus, on each subtask independently using task-specific linear classification heads with weighted BCE loss and dual-signal early stopping. Cross-task consistency is enforced post-hoc by nullifying type and stereotype labels for samples classified as non-misogynistic.

2. Related Work

Misogyny and Hate Speech Detection. Automated misogyny detection has grown as a subfield of hate speech detection following early benchmarks such as the Automatic Misogyny Identification (AMI) shared tasks at Evalita 2018 and 2020 [7, 8], which established binary misogyny detection and fine-grained category classification as standard evaluation protocols. Subsequent work demonstrated

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

arjunmukherjee.rs.cse23@itbhu.ac.in (A. Mukherjee); spal.cse@itbhu.ac.in (S. Pal)

0009-0009-4291-9625 (A. Mukherjee); 0000-0001-8743-9830 (S. Pal)

that transformer-based models consistently outperform feature-engineered and recurrent baselines on such tasks [9], a finding that motivates our own transformer-based approach.

Misogyny in Song Lyrics. Song lyrics present distinct challenges compared to social media text: longer documents, idiomatic and genre-specific language, high lexical density, and deliberate use of metaphor and euphemism to convey degradation. Early computational work on lyrics focused primarily on sentiment and emotion [10], with dedicated misogyny annotation efforts emerging only recently. These structural properties of lyrics — rather than the shared tasks themselves — inform our input representation choices, particularly the 256-token truncation limit and light normalisation strategy described in Section 4.

Spanish Pre-trained Language Models. BETO [5], a BERT-base model trained on a large Spanish corpus, established a strong monolingual Spanish baseline widely adopted across Spanish NLP tasks. XLM-RoBERTa [11] offers broader cross-lingual transfer but at the potential cost of language-specific representation quality, particularly for informal and culturally specific registers such as song lyrics. These trade-offs, examined empirically in Section 7, motivate our selection of BETO as the official submission model, given its strong performance across Spanish NLP tasks and its effective threshold-tuned fine-tuning in our experiments.

Multilabel Classification and Class Imbalance. Weighted loss functions, particularly Binary Cross-Entropy with positive class weights, are a widely adopted remedy for label imbalance in multilabel settings [12] and form the basis of our approach across all three subtasks. Per-label threshold calibration has been shown to yield substantial additional gains in such settings [13], and its absence in our current system is identified as a key limitation in Section 9.

Early Stopping in Low-Resource Fine-Tuning. With roughly 909 training samples available for the type and stereotype subtasks, transformer fine-tuning is prone to rapid memorisation. Standard patience-based early stopping on validation loss is a common safeguard [14] but can select poorly calibrated checkpoints when loss and task metric diverge. Our dual-signal mechanism — monitoring both macro-F1 stagnation and the val/train loss ratio simultaneously — is motivated by observations in low-resource NLP fine-tuning that loss-only stopping is insufficient in such regimes [15].

3. Task and Data Description

The MiSonGyny 2026 training corpus contains 2,303 Spanish-language song records balanced across dialects (1,181 `es_LATAM` and 1,122 `es_ES`). The test set comprises 576 unlabelled records.

Table 1 summarises the class distribution across all three subtasks.

Table 1
Class distribution in the MiSonGyny 2026 training corpus.

Subtask	Label	Count	Proportion
Subtask 1 — Detection	M — Misogynistic	1,009	43.8%
	NM — Non-misogynistic	1,294	56.2%
Subtask 2 — Type	Sexualization (S=1)	804	79.7% of M
	Violence (V=1)	314	31.1% of M
	Hate (H=1)	589	58.4% of M
Subtask 3 — Stereotype	Stereotypical (Y)	687	68.1% of M
	Non-stereotypical (N)	322	31.9% of M

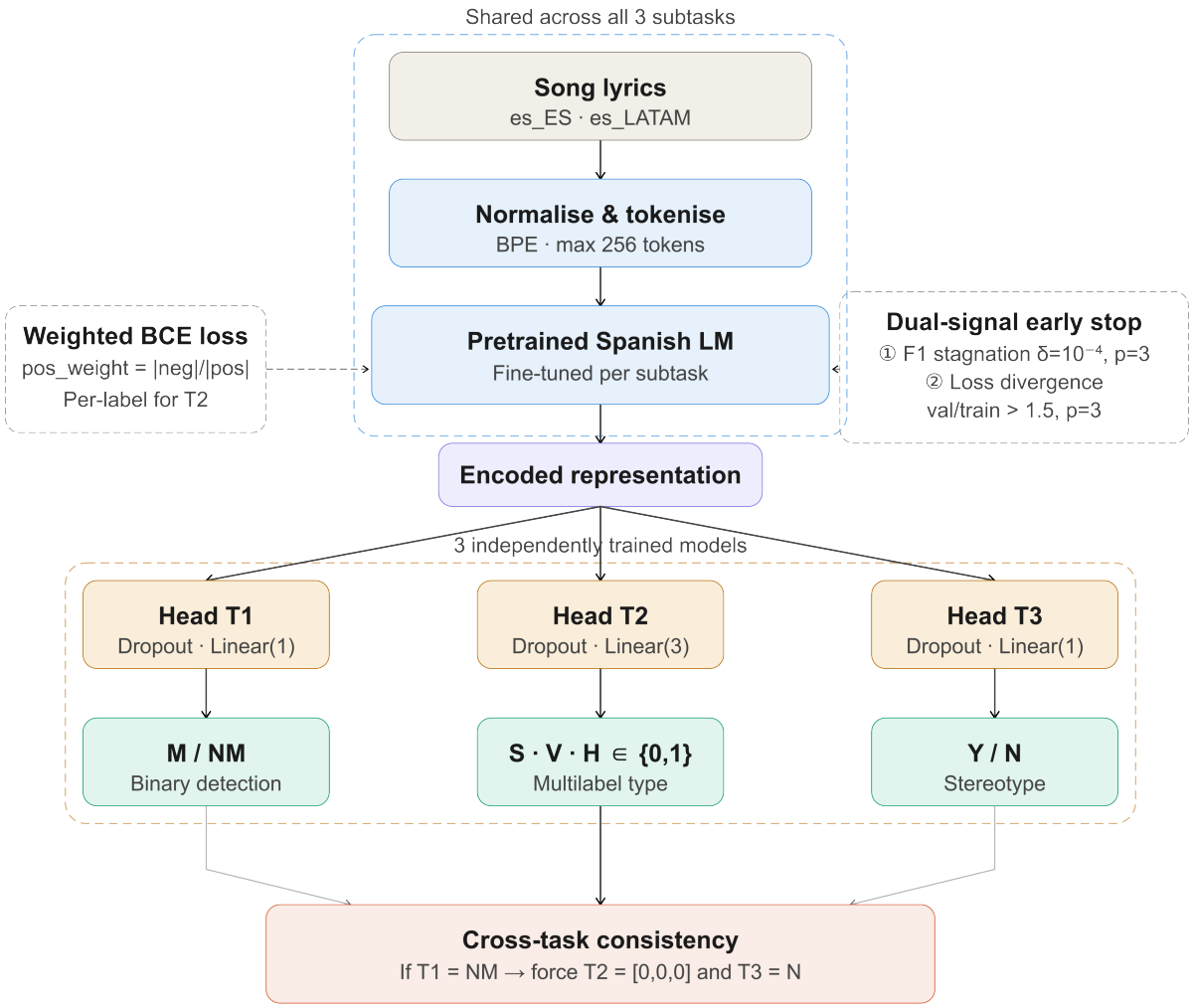

Figure 1: Overview of the methodological approach.

Annotation note: type and stereotype labels (Subtasks 2 & 3) are provided exclusively for M-labelled rows, making them semantically subordinate to Subtask 1. Submitted predictions follow `task_1_predictions.csv` (id, preds: M/NM), `task_2_predictions.csv` (id, S, V, H: 0/1), and `task_3_predictions.csv` (id, preds: Y/N), packaged as a single ZIP archive.

4. System Description

This section describes the overall system, shown in Figure 1.

4.1. Pre-trained Model

We build upon BETO (`dccuchile/bert-base-spanish-wwm-cased`) [5], a BERT-base architecture pre-trained on a large Spanish corpus using whole-word masking (WordPiece vocabulary: 31,002 tokens). Selection rationale: (i) strong Spanish monolingual pre-training; (ii) handles both Castilian and Latin American varieties; (iii) 12-layer transformer (hidden size 768, 12 heads, 110M parameters) with well-established performance across Spanish NLP classification tasks.

4.2. Input Representation

Light normalisation: escaped newlines replaced by whitespace, consecutive whitespace collapsed, Spanish characters (accents, ñ, ü) preserved. Lyrics are tokenised and truncated/padded to 256 tokens. Pilot analysis confirmed that >95% of training lyrics fit within this limit after normalisation, making truncation a minimal source of information loss.

4.3. Classification Head

A `BetoClassifier` wraps the encoder with a task-specific linear head. The `[CLS]` representation from the final layer passes through dropout ($p = 0.1$) and a linear projection. Output dimensionality by subtask:

- Subtask 1: 1 unit — binary sigmoid (M vs. NM)
- Subtask 2: 3 units — independent sigmoid per label (S, V, H)
- Subtask 3: 1 unit — binary sigmoid (Y vs. N)

Each subtask has its own independently trained model. No shared parameters across tasks, avoiding interference given that Subtasks 2 and 3 train on only ≈909 M-labelled rows.

4.4. Loss Function and Class Imbalance

All subtasks use **BCEWithLogitsLoss** with per-task positive class weights:

$$\text{pos\_weight} = \frac{|\text{negative class}|}{|\text{positive class}|} \tag{1}$$

For Subtask 2, weights are computed per label independently (S/V/H prevalences differ greatly: 79.7% / 31.1% / 58.4%). For Subtask 3, the weight upsamples the minority N class.

4.5. Optimisation

AdamW [16] with LR $= 2 \times 10^{-5}$, weight_decay $= 0.01$. Linear warmup over 10% of steps, then linear decay to zero. Gradient clip norm $= 1.0$. Maximum 8 epochs per subtask, subject to early stopping.

4.6. Dual-Signal Early Stopping

We apply a dual-signal early stopping mechanism:

- **F1 stagnation (primary):** Halt if validation macro-F1 fails to improve by $\delta = 1 \times 10^{-4}$ for a set number of consecutive epochs (patience), then restore the best-F1 checkpoint. Patience was set per subtask (1 for Subtask 1, 2 for Subtask 2, 3 for Subtask 3) to reflect their differing training-set sizes and convergence behaviour.
- **Loss divergence (secondary):** Halt if val_loss/train_loss > 1.5 for 3 consecutive epochs, even if F1 marginally improves — guards against poorly calibrated probabilities.

Weights are saved to CPU at every epoch of F1 improvement. The best-F1 checkpoint is restored before inference regardless of stopping reason.

4.7. Data Splits

15% held-out validation per subtask (1,957 train / 346 val for Subtask 1; 908 train / 101 val for Subtasks 2 & 3). Subtasks 1 & 3: stratified by label. Subtask 2: random split (multilabel). No external data or augmentation.

4.8. Cross-Task Prediction Consistency

Post-processing at inference enforces logical coherence:

- NM prediction (Subtask 1) $\Rightarrow$ forces $[\mathrm{S} = 0, \mathrm{V} = 0, \mathrm{H} = 0]$ in Subtask 2.
- NM prediction (Subtask 1) $\Rightarrow$ forces *preds*=N in Subtask 3.

5. Experimental Setup

Table 2 summarises all hyperparameters. All experiments were implemented using PyTorch and HuggingFace Transformers [17].

Table 2
Hyperparameters used in all experiments.

Hyperparameter	Value
Pre-trained model	`bert-base-spanish-wwm-cased` (BETO)
Max sequence length	256 tokens
Batch size	16
Maximum epochs	8 (subject to early stopping)
Early stopping patience	1 / 2 / 3 epochs (T1 / T2 / T3)
Early stopping δ (F1)	1×10^{-4} (macro-F1)
Loss divergence threshold	val_loss/train_loss > 1.5
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	10% of total steps
Gradient clip norm	1.0
Dropout (head)	0.1
Subtask 1 decision threshold	0.32 (per-epoch sweep 0.20–0.80)
Subtask 2 thresholds	per-label, tuned per epoch (best: S=0.77, V=0.80, H=0.40)
Validation split	15% (stratified T1 & T3; random T2)
Random seed	42
Framework	PyTorch + HuggingFace Transformers

6. Evaluation Metrics

All subtasks are evaluated using macro-averaged F1, Precision, and Recall (treating all classes equally regardless of frequency). Subtask 2 additionally reports:

- **Exact Match Ratio (EMR):** fraction of samples where the full predicted label set exactly matches gold. No partial credit.
- **Hamming Loss:** fraction of (sample, label) pairs where predicted $\neq$ gold. Lower is better; 0 is perfect.
- **ICM (Information Contrast Measure)** [18]: an information-theoretic similarity measure proposed by Amigó and Delgado for evaluating multi-label hierarchical extreme classification. ICM generalises pointwise mutual information to compare a system prediction against the gold label set: it rewards the information shared between the two label sets and penalises the

information each set contributes on its own, with the relative weight of these terms controlled by fixed coefficients. Because it is defined over label sets rather than per-label decisions, it accounts for the relative specificity and frequency of labels rather than treating all errors equally as Hamming Loss does. We report the value produced by the official MiSonGyny 2026 shared-task scorer.

7. Results

The following results are organised to reflect the two stages of our experimental pipeline: exploratory runs with XLM-RoBERTa and BERTIN that informed the final system design, and the official BETO submission — which achieved the best validation performance — evaluated against the held-out test set at MiSonGyny 2026.

7.1. Cross-Model Results Summary (Subtask 1)

Table 3 compares best validation results across all model iterations on Subtask 1 (binary misogyny detection). The BETO system is the official submission, selected on the basis of achieving the highest validation F1 (0.7983) with per-epoch threshold tuning. XLM-R and BERTIN results are from prior exploratory runs. All runs use an 85/15 split (1,957 train / 346 val) on the 2,303-record corpus.

Table 3
Subtask 1 cross-model comparison (validation set, macro-averaged). ★ = official submission (BETO).

Model	Run	Best Ep.	Macro F1	Thr.	ES
XLM-RoBERTa	Run 1	3 / 4	0.7882	argmax	No
XLM-RoBERTa	Run 2	1 / 8	0.7894	argmax	Yes (p=1)
XLM-RoBERTa	Run 3	1 / 8	0.7919	0.56	Yes (p=1)
BETO ★	Run 4	3 / 8	0.7983	0.32	Yes (p=1)
BERTIN	Expl.	5 / 18	0.7480	0.5	Yes (p=3)

7.2. Full Results — All Three Subtasks (BETO Official System)

Table 4 reports BETO validation results across all three subtasks (T1: 1,957 train / 346 val; T2 and T3: 908 train / 101 val).

Table 4
BETO official system — validation results (macro-averaged).

Subtask	Best Ep.	Macro F1	Macro P	Macro R
T1 — Detection	3 / 8	0.7983	0.7975	0.7994
T2 — Type (S/V/H)	3 / 5	0.6599	0.6167	0.7324
T3 — Stereotype	4 / 8	0.6808	0.6765	0.6893

T1 stopped at epoch 4 (no F1 improvement, patience$=1$). T2 stopped at epoch 5 (no F1 improvement for 2 epochs, patience$=2$). T3 stopped at epoch 7 (F1 stagnation 3/3 + ratio $=19.25$, 3/3).

Test-set predictions: T1 $\rightarrow$ 218 M / 358 NM. T2 $\rightarrow$ S=170, V=132, H=183. Official test-set scores are reported in Section 8.

7.3. BETO Epoch-Level Training Logs

7.3.1. Subtask 1 — Binary Misogyny Detection

Train: 1,957 | Val: 346 | Best epoch: 3 | Best Macro F1: 0.7983 | Threshold: 0.32

Table 5
BETO Subtask 1 epoch log (official submission). ★ = best F1 checkpoint; threshold tuned per epoch.

Ep	Tr.L	Val L	F1	P	R	Thr	Notes
1	0.7208	0.8987	0.7598	0.7815	0.7558	0.68	Best
2	0.7159	1.0266	0.7845	0.7950	0.7810	0.31	Best
3	0.5979	0.9349	0.7983	0.7975	0.7994	0.32	Best ★
4	0.4623	1.2784	0.7977	0.8096	0.8096	0.66	**STOP**

Observation: BETO achieves its best validation F1 (0.7983) at epoch 3 with a tuned threshold of 0.32, then stops at epoch 4 as F1 fails to improve beyond patience$= 1$. The per-epoch threshold sweep (0.20–0.80) provides a clear advantage over a fixed 0.5 threshold.

7.3.2. Subtask 2 — Type Classification (S / V / H)

Train: 908 (M-only) | Val: 101 | Best epoch: 3 | Best Macro F1: 0.6599

Table 6
BETO Subtask 2 epoch log. Per-label decision thresholds were tuned each epoch; the best checkpoint (epoch 3) used S=0.77, V=0.80, H=0.40.

Ep	Tr.L	Val L	F1	P	R	Notes
1	0.9373	0.9756	0.5932	0.5093	0.7427	Best
2	0.7391	1.0191	0.6379	0.6380	0.6437	Best
3	0.5317	0.8661	0.6599	0.6167	0.7324	Best ★
4	0.3227	1.3116	0.6414	0.6234	0.6661	No imp. (1/2)
5	0.1661	1.7755	0.6247	0.6626	0.5918	No imp. (2/2) **STOP**

Observation: Best F1=0.6599 at epoch 3 with per-label tuned thresholds (S=0.77, V=0.80, H=0.40), favouring recall (0.7324) over precision (0.6167). Validation loss rises steadily from epoch 3 onward (0.87→1.31→1.78) while train loss falls to 0.17, indicating rapid overfitting on the small effective signal for the minority labels. Early stopping (patience 2 on macro-F1) halts at epoch 5. A per-label breakdown of this checkpoint is given in Table 10 (Section 9).

7.3.3. Subtask 3 — Stereotype Identification

Train: 908 (M-only) | Val: 101 | Best epoch: 4 | Best Macro F1: 0.6808

Table 7
BETO Subtask 3 epoch log.

Ep	Tr.L	Val L	F1	P	R	Notes
1	0.4433	0.4275	0.6140	0.7155	0.6116	Best
2	0.4469	0.4248	0.4059	0.3416	0.5000	No imp. (1/3)
3	0.4328	0.3913	0.6154	0.6216	0.6397	Best
4	0.3339	0.3770	0.6808	0.6765	0.6893	Best ★
5	0.1921	0.4810	0.5640	0.6331	0.6393	(1/3) ratio=2.50
6	0.1114	1.3893	0.5710	0.7903	0.5865	(2/3) ratio=12.47
7	0.0414	0.7969	0.6747	0.6752	0.7000	ratio=19.25 **STOP**

Observation: Severe overfitting — by epoch 7, train loss $= 0.04$ vs val loss $= 0.80$ (ratio$= 19.25$). The model memorises training data quickly given the small set. Best F1$= 0.6808$ at epoch 4 achieves the most balanced precision/recall (0.68/0.69). Dual-signal stopping correctly halts at epoch 7 before further degradation.

7.4. Exploratory Models — Subtask 1

Prior to the BETO submission we ran three exploratory XLM-RoBERTa configurations and one BERTIN run on Subtask 1, all using an 85/15 split on 2,303 records with chunking (MAX_LEN=384, STRIDE=128) and class-weighted CrossEntropyLoss; per-epoch threshold tuning was applied to XLM-R Run 3 only. All four converged within their first few epochs: XLM-R Runs 1–3 peaked at epochs 3, 1, and 1 (best macro-F1 0.7882, 0.7894, and 0.7919 respectively), while BERTIN peaked later at epoch 5 (0.7480). Best-checkpoint figures for each run are reported in the cross-model summary (Table 3), and their per-epoch macro-F1 trajectories appear in Table 8; we omit the full epoch-level logs for these abandoned runs for brevity. The BETO official submission logs are given in Section 7 above.

7.5. Macro-F1 Progression Across All Models (Subtask 1)

Table 8 consolidates the per-epoch validation macro-F1 trajectories of all five Subtask 1 configurations into a single view, making the convergence behaviour directly comparable. Two patterns stand out. First, the XLM-RoBERTa runs (R1–R3) and BETO all reach their best checkpoint within the first three epochs and then stop early, whereas BERTIN converges more slowly, peaking only at epoch 5. Second, BETO attains the highest macro-F1 of any configuration (0.7983 at epoch 3), which is the basis for its selection as the official submission. Dashes denote epochs not run (early-stopped or beyond a fixed budget); "STOP" marks the epoch at which early stopping halted training.

Table 8
Macro-F1 per epoch across all Subtask 1 models. ★ = official submission (BETO, epoch 3).

Ep	R1	R2	R3	BETO	BERTIN
1	0.7792	0.7894	0.7919	0.7598	0.7108
2	—	STOP	STOP	0.7845	0.7248
3	0.7882	—	—	0.7983★	0.7462
4	0.7860	—	—	STOP	0.7078
5	—	—	—	—	0.7480
6	—	—	—	—	STOP

R1/R2/R3 = XLM-RoBERTa Runs 1–3; BETO (★, official submission) and R3 with threshold tuning; BERTIN exploratory.

8. Official Competition Results

Table 9 reports the official test-set rankings released by the MiSonGyny 2026 organisers across all three subtasks. Our system (IReL_IIT(BHU)) is submitted under the CodaBench username `arjun108`.

Our system ranks **7th out of 12** on Task 1 (F1= 0.8434), **6th out of 12** on Task 2 (F1= 0.6737), and **11th out of 12** on Task 3 (F1= 0.6366). The official test F1 on Task 1 (0.8434) exceeds our validation estimate (0.7983), suggesting the BETO model generalises well to the held-out test distribution and that the validation set, while representative, was somewhat harder than the test set. Task 2 performance is close to the validation estimate (0.6737 vs. 0.6599), confirming that the dual-signal early stopping generalises well on that subtask. Task 3 shows a more pronounced drop (0.6366 vs. 0.6808), consistent with the severe overfitting observed in the epoch logs (train/val loss ratio reaching 19.25 at stopping), suggesting the checkpoint selected on the small validation set did not fully generalise to the held-out test distribution.

9. Discussion

The results presented in Section 7 address three inter-related questions: why BETO was selected as the official submission over other tested models, what drives the high recall but low precision pattern observed in Subtask 2, and why Subtask 3 shows the most pronounced gap between validation and official test performance — each of which we address in turn below.

Table 9
Official MiSonGyny 2026 final rankings (macro-averaged F1, best submission per participant). **Bold** = our system (`arjun108`).

Task 1 — Detection			Task 2 — Type			Task 3 — Stereotype		
Rank	**Team**	**F1**	**Rank**	**Team**	**F1**	**Rank**	**Team**	**F1**
1	tbernal	0.8856	1	tbernal	0.7385	1	tbernal	0.8283
2	ikanate	0.8712	2	manuelsanz	0.7079	2	ikanate	0.8109
3	aerojash	0.8581	3	vramos	0.6939	3	aerojash	0.8071
4	manuelsanz	0.8559	4	cicese-lcdaa	0.6846	4	franrodrigo	0.8039
4	vramos	0.8559	5	ikanate	0.6813	5	manuelsanz	0.8015
5	cicese-lcdaa	0.8533	**6**	**arjun108**	**0.6737**	5	vramos	0.8015
6	franrodrigo	0.8448	7	aerojash	0.6699	6	cicese-lcdaa	0.7924
7	**arjun108**	**0.8434**	8	amisadai_ed	0.6683	7	amisadai_ed	0.7860
8	VerbaNexAI Lab	0.8417	9	franrodrigo	0.6645	8	VerbaNexAI Lab	0.7798
9	sergiomarinnx	0.8334	10	sergiomarinnx	0.6562	9	sergiomarinnx	0.7724
10	amisadai_ed	0.8219	11	misongyny	0.5015	10	misongyny	0.7206
11	misongyny	0.7931	12	VerbaNexAI Lab	0.4022	**11**	**arjun108**	**0.6366**

9.1. Strengths

BETO's large-scale Spanish pre-training provides a strong linguistic foundation for both `es_ES` and `es_LATAM` lyrics. Per-class weighted BCE loss handles imbalance in all three subtasks without oversampling. The dual-signal early stopping mechanism provides robust overfitting protection critical for Subtasks 2 and 3 (≈909 training samples). Per-epoch threshold tuning on Subtask 1 further improves discrimination. Cross-task post-processing enforces logical coherence, reducing contradictory predictions across evaluation metrics.

9.2. Analysis of Cross-Model Results

On Subtask 1, BETO achieves the highest validation F1 (0.7983) among all tested configurations, justifying its selection as the official submission. Three factors explain its advantage over the next-best models (XLM-R Run 3: 0.7919, BERTIN exploratory: 0.7480):

- **Per-epoch threshold tuning:** BETO applies a sweep (0.20–0.80) at each epoch, finding an optimal threshold of 0.32 at epoch 3. This provides a meaningful gain over a fixed 0.5 threshold as used in the BERTIN exploratory run.
- **Larger validation set:** The 85/15 split (346 val samples) produces more reliable F1 estimates than BERTIN's 90/10 split (231 val samples), making model selection more trustworthy.
- **Efficient convergence:** BETO reaches its peak in only 3 epochs, suggesting rapid adaptation of its Spanish BERT representations to the misogyny detection task, with early stopping correctly halting at epoch 4.

9.3. Subtask 2 & 3 Analysis

Subtask 2 (type classification) is the hardest task with F1= 0.6599. The Violence label (31.1% prevalence) is the primary challenge — its underrepresentation yields the weakest per-label F1 (0.5085) despite per-label threshold tuning, with precision (0.4225) lagging recall (0.6383). Table 10 breaks down validation performance by label at the best checkpoint (epoch 3) and confirms that the macro-F1 figure masks substantial per-label variation: Violence trails both Sexualization and Hate by a wide margin, consistent with its lower prevalence in the training data.

Subtask 3 (stereotype detection) achieves F1= 0.6808. The rapid overfitting (train loss 0.04 by epoch 7) despite a relatively balanced label distribution (Y= 68%, N= 32%) suggests the small training set (≈909 samples) is the primary bottleneck. The dual-signal early stopping correctly identifies and halts the divergence (ratio= 19.25 at epoch 7).

Table 10
Subtask 2 — per-label validation results at the best checkpoint (epoch 3). Prevalence is reported as percentage of M-labelled training songs; per-label thresholds at this checkpoint were S=0.77, V=0.80, H=0.40.

Label	Prev.	F1	P	R
Sexualization (S)	79.7%	0.7982	0.8505	0.7521
Violence (V)	31.1%	0.5085	0.4225	0.6383
Hate (H)	58.4%	0.6730	0.5772	0.8068
Macro avg.	—	**0.6599**	**0.6167**	**0.7324**

9.4. Limitations and Future Work

- Truncation to 256 tokens — sliding-window encoding could better handle longer songs with less information loss.
- No multi-task learning — a joint model could leverage shared representations across all three subtasks.
- Per-label thresholds for Subtask 2 were tuned on a single validation split — the high optimal thresholds (S=0.77, V=0.80, H=0.40) may overfit the validation set, and cross-validated calibration could yield more robust thresholds.
- No dialect conditioning — explicit dialect-aware representations could improve `es_LATAM` vs. `es_ES` performance.
- Small training set for T2/T3 — external Spanish misogyny data or augmentation strategies could help.

10. Conclusion

We presented the IReL_IIT(BHU) system for MiSonGyny 2026 at IberLEF 2026, addressing automatic misogyny detection in Spanish song lyrics across all three subtasks. Our approach fine-tunes BETO, a Spanish BERT-base model, with task-specific linear heads, weighted BCE loss, per-epoch threshold tuning, and dual-signal early stopping monitoring both F1 stagnation and loss divergence. Cross-task consistency is enforced via post-processing at inference time. The official BETO submission achieves validation Macro F1 of 0.7983 / 0.6599 / 0.6808 across the three subtasks, ranking 7th on Task 1, 6th on Task 2, and 11th on Task 3 in the official competition. Future work will explore multi-task learning, per-label threshold optimisation for Subtask 2, and dialect-aware encoding.

Acknowledgements

The authors gratefully acknowledge Krishna Tewari for her valuable insights and constructive guidance at every stage of this work, from participation through to final submission.

Declaration on Generative AI

During the preparation of this work, the authors used AI-assisted tools to support code debugging, manuscript drafting, and submission formatting. All content was reviewed and edited by the authors, who take full responsibility for the publication.

References

[1] 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 (2026).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[3] T. Alcántara, M. Soto, C. Macías, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveroón, Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural 75 (2025). URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln.

[4] H. Gómez-Adorno, G. Bel-Enguix, G. Sierra, S.-T. Andersen, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macías, 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). URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-trained BERT Model and Evaluation Data, in: Practical ML for Developing Countries Workshop, ICLR 2020, 2020. URL: https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased.

[6] 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[7] E. Fersini, P. Rosso, 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), volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 214–228. URL: https://ceur-ws.org/Vol-2150/.

[8] E. Fersini, D. Nozza, P. Rosso, AMI @ Evalita 2020: Automatic Misogyny Identification, in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/.

[9] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, in: Complex Networks and Their Applications VIII, Springer, 2020, pp. 928–940. doi:10.1007/978-3-030-36687-2_77.

[10] M. Fell, C. Sporleder, Lyrics-Based Analysis and Classification of Music, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, 2014, pp. 620–631. URL: https://aclanthology.org/C14-1059.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.

[13] R.-E. Fan, C.-J. Lin, A Study on Threshold Selection for Multi-Label Classification, Technical Report, National Taiwan University, 2007. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/threshold.pdf.

[14] L. Prechelt, Early Stopping — But When?, in: Neural Networks: Tricks of the Trade, Springer, 1998, pp. 55–69. doi:10.1007/3-540-49430-8_3.

[15] M. Mosbach, M. Andriushchenko, D. Klakow, On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines, in: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), 2021. URL: https://openreview.net/forum?id=nzpLWnVAyah.

[16] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: 7th International Conference on Learning Representations, ICLR 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[17] 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 (EMNLP 2020), Association for Computational Linguistics, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[18] E. Amigó, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.