VerbaNex AI Lab at MiSonGyny-IberLEF 2026: BETO-Based Detection and Classification of Misogynistic Content in Mexican and Spanish Language Song Lyrics

Harold Lagares[1], Edwin Puertas[1], Juan Carlos Martinez-Santos[1] and Jairo Serrano[1]

[1]*Escuela de Transformación Digital, Universidad Tecnológica de Bolívar, Cartagena, Colombia*

Abstract

This paper describes the system submitted by team VerbaNex AI Lab to the MiSonGyny 2026 shared task at IberLEF 2026, which focuses on the detection and classification of misogynistic content in Spanish-language song lyrics across three subtasks. Task 1 requires binary detection of misogyny (M / NM) over the full corpus; Task 2 requires fine-grained classification of the type of misogyny (Hate, Sexualization, or Violence) exclusively for misogynistic songs; and Task 3 requires binary detection of gender stereotypes (Y / N), also restricted to misogynistic songs. Our approach progresses from a classical machine learning baseline using TF-IDF representations and Support Vector Machines to Transformer-based systems fine-tuned on BETO (`dccuchile/bert-base-spanish-wwm-cased`), a monolingual Spanish BERT model. For Task 2, the severe class imbalance of the Violence category (only 4% of misogynistic songs) motivated a two-stage cascade architecture: a specialized BETO binary classifier for Violence detection with oversampling, followed by a second BETO model for Hate vs. Sexualization discrimination. Our final system achieves Macro F1-scores of 0.8417 on Task 1, 0.4022 on Task 2, and 0.7798 on Task 3, demonstrating the effectiveness of monolingual Spanish Transformers for this challenging domain-specific NLP task.

Keywords

MiSonGyny, misogyny detection, Spanish NLP, BETO, song lyrics, text classification, IberLEF 2026

1. Introduction

The automatic detection of misogynistic content in online and cultural media has become an increasingly important task in Natural Language Processing (NLP), as misogyny in music contributes to the normalization of gender-based discrimination and violence [1]. Spanish-language song lyrics represent a particularly challenging domain given the prevalence of misogynistic tropes in popular genres such as reggaeton, corridos, and urban trap, combined with the figurative and idiomatic nature of lyrical language. The MiSonGyny 2026 shared task, organized within the IberLEF 2026 evaluation campaign [2], addresses this challenge with a hierarchical annotation scheme that captures not only whether a song is misogynistic but also the type of misogyny it conveys and whether it reinforces gender stereotypes.

MiSonGyny 2026 builds upon prior related evaluations. The MiSonGyny 2025 task [3] established baselines for misogyny detection in this domain, and the HOMO-MEX 2024 task [4] addressed hate speech detection towards the Mexican Spanish-speaking LGBT+ population in song lyrics, serving as a direct predecessor.

The MiSonGyny 2026 task [5] comprises three subtasks evaluated with Macro F1-score: binary misogyny detection (Task 1), fine-grained type classification (Task 2), and gender stereotype detection (Task 3). A key structural feature of the task is that Task 2 and Task 3 apply only to songs already identified as misogynistic in Task 1, creating a hierarchical prediction pipeline.

The VerbaNex AI Lab at the Universidad Tecnológica de Bolívar has an active track record in NLP shared tasks involving Transformer-based models. Particularly relevant to this work are our participation in multi-label scientific discourse detection using DeBERTa at CheckThat! 2025 [6] and

IberLEF 2026, September 2026, León, Spain

hlagares@utb.edu.co (H. Lagares); epuerta@utb.edu.co (E. Puertas); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); jserrano@utb.edu.co (J. Serrano)

0009-0007-6693-7188 (H. Lagares); 0000-0002-0758-1851 (E. Puertas); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0001-8165-7343 (J. Serrano)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-09-29

CEUR-WS.org/Vol-4273/MiSonGyny2026_paper4.pdf

emotion classification in Spanish with RoBERTa at SemEval-2025 [7]. This experience with fine-tuning under class imbalance directly motivated our approach to MiSonGyny 2026.

The primary challenges of this task are: (1) the idiomatic and figurative nature of song lyrics, where misogyny is often implicit; (2) the cascaded task structure, where Task 2 and Task 3 errors compound with Task 1 errors; (3) extreme class imbalance in Task 2, with Violence comprising only 4% of misogynistic songs; and (4) the diversity of Spanish dialects and slang across musical genres. We address these with a pipeline combining domain-adapted preprocessing, class imbalance strategies, and fine-tuning of BETO [8], a large-scale monolingual Spanish BERT model.

The rest of this paper is structured as follows. Section 2 describes the task and data. Section 3 presents our system. Section 4 reports experimental results. Section 5 concludes the paper.

2. Task and Dataset Description

2.1. Task Overview

The MiSonGyny 2026 task [5] provides a corpus of Spanish song lyrics with hierarchical annotations across three interdependent subtasks.

Task 1 (Binary Misogyny Detection) requires classifying each song as Misogynistic (M) or Non-Misogynistic (NM). The evaluation metric is Macro F1-score.

Task 2 (Misogyny Type Classification) applies exclusively to songs classified as M and requires identifying the type of misogynistic content among three mutually exclusive categories: Hate (H), Sexualization (S), and Violence (V). The evaluation metric is Macro F1-score.

Task 3 (Gender Stereotype Detection) also applies exclusively to misogynistic songs and requires determining whether the song contains gender stereotypes (Y) or not (N). The evaluation metric is Macro F1-score.

2.2. Dataset

The dataset consists of 2,303 annotated Spanish song lyrics drawn from multiple genres including reggaeton, trap, and urban corridos. Table 1 summarises the dataset statistics and class distributions for each task.

Table 1
MiSonGyny 2026 dataset statistics and class distribution.

Task	Class	Count	%	Scope
Task 1	NM (Non-Misogynistic)	1,294	56.2	Full corpus
	M (Misogynistic)	1,009	43.8	
Task 2	H (Hate)	589	58.4	M songs only
	S (Sexualization)	380	37.7	
	V (Violence)	40	4.0	
Task 3	Y (Stereotype)	687	68.1	M songs only
	N (No Stereotype)	322	31.9	

Task 2 presents the most severe class imbalance: the Violence class comprises only 4.0% of misogynistic songs (40 samples), making it extremely difficult to learn a reliable decision boundary. The hierarchical task structure means that Task 2 and Task 3 training and inference are restricted to the 1,009 songs labeled as misogynistic in Task 1.

An additional challenge is text length: 31.35% of songs exceed approximately 400 words (note: words and tokens are not equivalent; one word typically tokenizes into 1–3 WordPiece subword tokens, so 400 words correspond roughly to 500–600 tokens, approaching the 512-token hard limit of BERT-based models). The mean song length is 348 words (std: 203), with a maximum of 2,258 words.

3. System Description

Our approach evolves across two tiers of increasing complexity, mirroring the progression strategy adopted in other VerbaNex AI Lab submissions [6]: a classical machine learning baseline followed by fine-tuned Transformer models. Figure 1 summarises the full system pipeline.

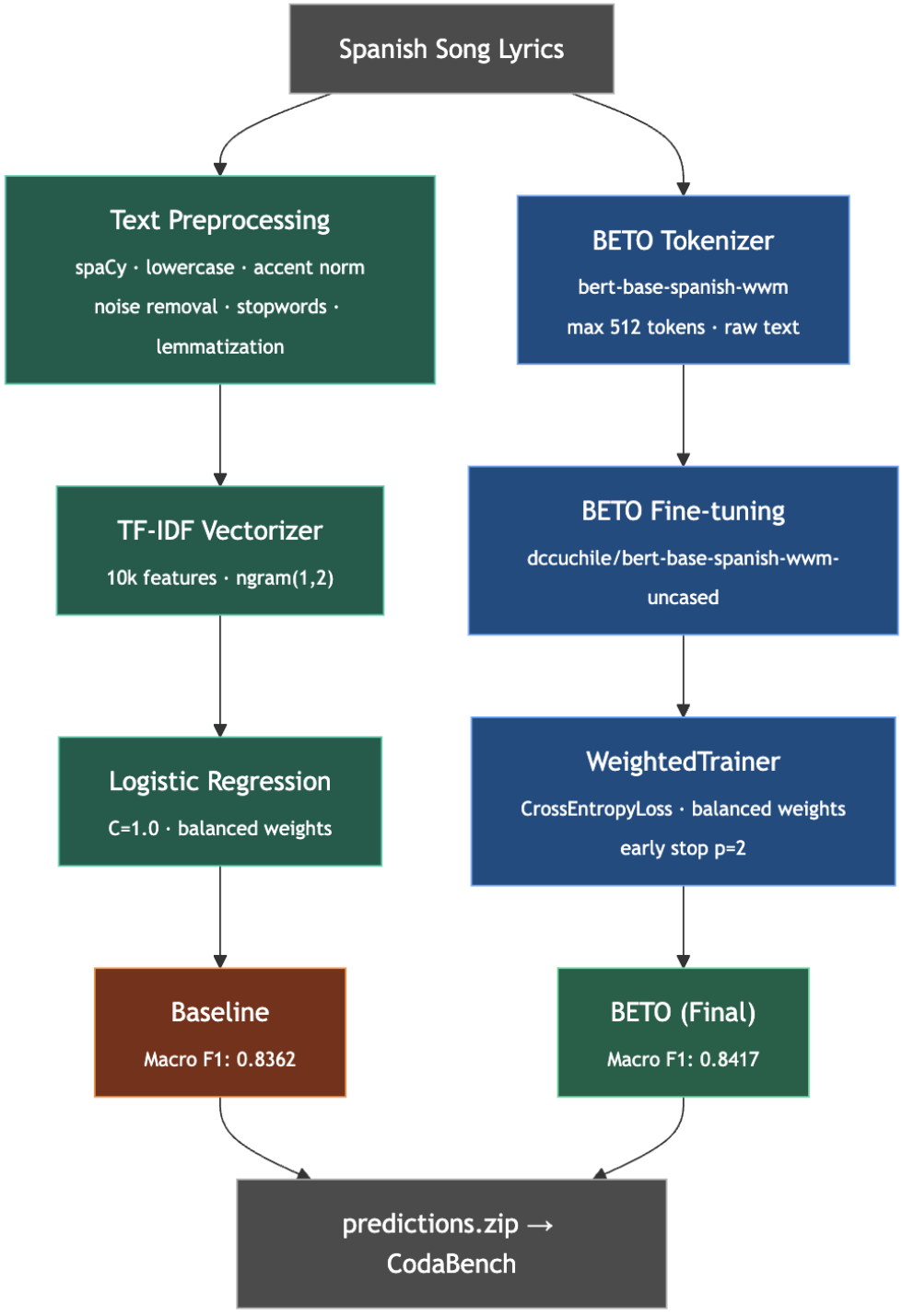

Figure 1: Full system pipeline for MiSonGyny 2026. The baseline path (left) uses TF-IDF features with SVM for all three tasks. The main system path (right) fine-tunes the cased BETO model (`dccuchile/bert-base-spanish-wwm-cased`) with task-specific architectures: standard fine-tuning for Tasks 1 and 3, and a two-stage cascade for Task 2.

3.1. Text Preprocessing

We apply a uniform preprocessing pipeline to the baseline path:

1. **Lowercasing**: All text converted to lowercase.
2. **HTML and noise removal**: HTML tags, line breaks, and multiple spaces removed and normalized.
3. **Special character filtering**: Punctuation and digits removed while preserving Spanish-specific characters (ñ, tildes).

For the BETO-based models, raw text (without preprocessing) is fed directly to the BERT tokenizer, following standard fine-tuning practice for pre-trained language models [8]. Songs exceeding 512 tokens are truncated, affecting approximately 31% of the corpus.

3.2. Baseline: TF-IDF + SVM

We train a Support Vector Machine on TF-IDF features with `max_features=10000`, `ngram_range=(1,3)`, `max_df=0.8`, and `min_df=2`, using a linear kernel. To counteract class imbalance, per-class weights are computed via `compute_class_weight('balanced')` from scikit-learn [9]. The same pipeline is applied independently to all three tasks, with Task 2 and Task 3 trained and evaluated exclusively on misogynistic songs. Local Macro F1-scores on a 80/20 stratified split are: 0.7573 (Task 1), 0.4325 (Task 2), and 0.5930 (Task 3).

3.3. Main System: BETO Fine-tuning

Our primary system fine-tunes BETO (`dccuchile/bert-base-spanish-wwm-cased`) [8], a Spanish BERT model pre-trained with whole-word masking on a large Spanish corpus. BETO was selected over multilingual models due to its superior performance on Spanish-specific NLP benchmarks. Figure 2 illustrates the task-specific architectures.

3.3.1. Tasks 1 and 3: Binary Classification

For Task 1 (full corpus) and Task 3 (misogynistic songs only), BETO is fine-tuned for standard 2-class sequence classification using a custom `WeightedTrainer` that subclasses Hugging Face `Trainer` [9] and injects class-balanced weights via `compute_class_weight('balanced')` into the CrossEntropyLoss:

```
class WeightedTrainer(Trainer):
  def compute_loss(self, model, inputs,
                   return_outputs=False, **kwargs):
    labels = inputs.pop("labels")
    outputs = model(**inputs)
    logits = outputs.logits
    loss_fct = nn.CrossEntropyLoss(
        weight=class_weights_tensor)
    loss = loss_fct(
        logits.view(-1,
            self.model.config.num_labels),
        labels.view(-1))
    return (loss, outputs) \
        if return_outputs else loss
```

Task 1 hyperparameters: learning rate 1×10^{-5}, batch size 8, gradient accumulation steps 2, 5 epochs, weight decay 0.01, max sequence length 512 tokens, Early Stopping (patience 2).

Task 3 hyperparameters: learning rate 2×10^{-5}, batch size 8, 3 epochs, weight decay 0.01, max sequence length 512 tokens, Early Stopping (patience 2).

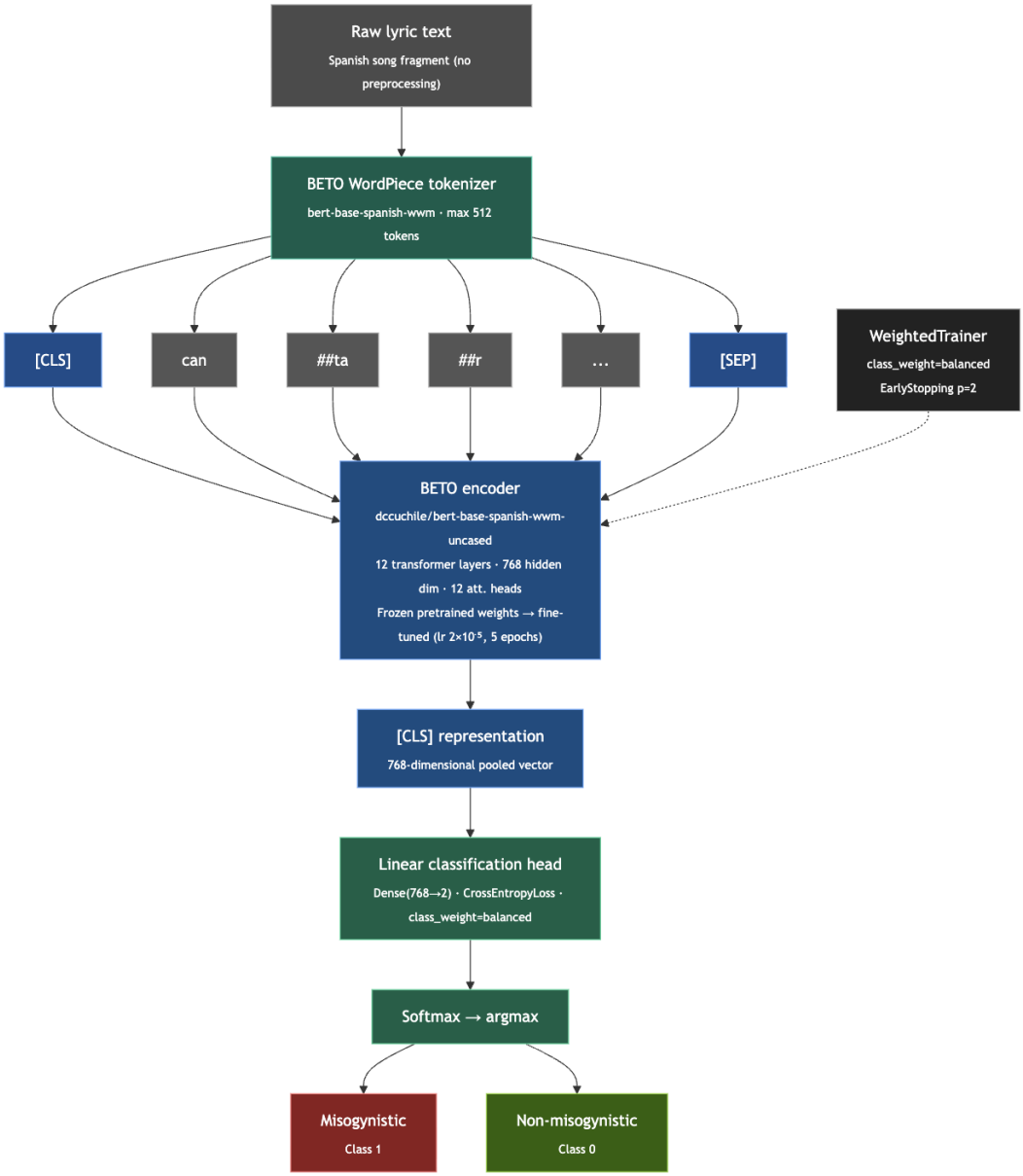

Figure 2: BETO fine-tuning architectures. Left: Tasks 1 and 3 use a standard binary classification head with `WeightedTrainer` and CrossEntropyLoss. Right: Task 2 uses a two-stage cascade — Model A detects Violence vs. Non-Violence (with oversampling); Model B discriminates Hate vs. Sexualization among Non-Violence predictions. All models used the cased BETO variant (`dccuchile/bert-base-spanish-wwm-cased`).

3.3.2. Task 2: Two-Stage Cascade for Type Classification

Task 2 presents an extreme imbalance challenge: Violence (class V) comprises only 40 training samples (4.0%) among the 1,009 misogynistic songs, making direct 3-class classification unreliable. After experimenting with a single 3-class `WeightedTrainer` (local Macro F1: 0.5136 on validation, but degrading on the test set), we designed a **two-stage cascade** architecture:

Model A — Violence vs. Non-Violence: A binary BETO classifier trained exclusively to detect Violence. To compensate for the severe imbalance (32 Violence vs. 775 Non-Violence training instances), we apply random oversampling of the Violence class to reach a 1:4 ratio (193 Violence, 775 Non-Violence). Class weights are set to Non-Violence: 0.625, Violence: 5.016.

Model B — Hate vs. Sexualization: A binary BETO classifier trained on the subset of non-violent misogynistic songs (471 Hate, 304 Sexualization in training). Class weights: Hate 0.823, Sexualization 1.275.

Cascade inference: A song first passes through Model A. The Violence probability threshold θ was selected via a grid search over the range $[0.15, 0.70]$ with step 0.05 on the validation split, optimizing the Macro F1-score of the full three-class system. The optimal value $\theta = 0.60$ maximized the validation Macro F1-score while maintaining acceptable Violence precision. If the Violence probability exceeds θ, the song is classified as Violence; otherwise it is passed to Model B for Hate vs. Sexualization discrimination. Both Model A and Model B share the same training configuration: learning rate 2×10^{-5}, batch size 8, 5 epochs, weight decay 0.01, max sequence length 512 tokens, Early Stopping (patience 2).

4. Experiments and Results

We report results across two stages: the CodaBench development phase (four submission rounds) and the official final testing phase.

4.1. Development Phase

Table 2 shows the Macro F1-scores on the CodaBench development phase across our four submission rounds. Submission 1 uses the SVM baseline for all tasks. Submissions 2 and 3 explore intermediate BETO configurations for Tasks 1 and 3 and a single-model 3-class BETO for Task 2. Submission 4 is the final system with the two-stage cascade for Task 2.

Table 2
Results on the CodaBench development phase (Macro F1-score).

Submission	Task 1	Task 2	Task 3
Submission 1 (SVM baseline)	0.8362	0.3967	0.7595
Submission 2 (BETO)	0.8417	0.4084	0.7798
Submission 3 (BETO)	0.8362	0.4271	0.6836
Submission 4 (BETO + cascade)	0.8417	0.4022	0.7798

Across all submissions, BETO consistently outperforms the SVM baseline on Tasks 1 and 3. Task 1 improves from 0.8362 (baseline) to 0.8417 (BETO). Task 3 improves from 0.7595 (baseline) to 0.7798 (BETO). Task 2 remains the most challenging: the two-stage cascade (Sub. 4, F1: 0.4022) achieves a comparable result to the best intermediate BETO configuration (Sub. 3, F1: 0.4271) on the development set. Notably, the baseline SVM obtains precision=0.0 on Task 2 in Submission 1, indicating total failure on the Violence class, which the cascade system partially addresses.

4.2. Final Testing Phase

Table 3 presents the official results on the testing phase. Per the organizers' policy, the highest Macro F1-score across all submissions was used to determine each team's official position.

Table 3
Official results on the final testing phase (best submission per task).

Task	Macro F1	Precision	Recall
Task 1 (Binary misogyny)	0.8417	0.8409	0.8426
Task 2 (Type classification)	0.4022	0.5537	0.3741
Task 3 (Stereotypes)	0.7798	0.7723	0.7904

4.3. Discussion

Tasks 1 and 3 benefit substantially from BETO's contextual representations. For Task 1, the improvement over the SVM baseline (0.8362 to 0.8417) confirms that Transformer encoders capture implicit misogynistic cues — such as figurative language and slang — that n-gram TF-IDF features miss. Task 3 shows a similar pattern (0.7595 to 0.7798), suggesting that gender stereotype signals are distributed across lyrical context in ways that exceed the reach of bag-of-words models.

Task 2 remains the central challenge of this shared task. The Violence class (40 training samples) is too sparse for any standard fine-tuning approach to learn reliably. Despite oversampling and class weighting in Model A, the cascade system achieves a Violence recall of only 0.1250 on the development set. The precision-recall asymmetry on the test set (Precision: 0.5537, Recall: 0.3741) suggests the system over-predicts Hate at the expense of Violence and Sexualization. Future work should explore retrieval-augmented augmentation of the Violence class, adversarial data augmentation, or few-shot learning approaches to address this structural bottleneck.

The performance gap between Tasks 1 and 3 (strong results) and Task 2 (limited results) is consistent with the general finding from VerbaNex AI Lab's experience with classification tasks involving extreme class imbalance [6, 7]: Transformer-based encoders substantially improve performance on well-represented classes but struggle with minority classes where the volume of training signal is insufficient to overcome the pre-training distribution.

5. Conclusion

We presented the VerbaNex AI Lab system for the MiSonGyny 2026 shared task at IberLEF 2026. Our approach fine-tunes BETO (`dccuchile/bert-base-spanish-wwm-cased`) across three subtasks: standard binary classification with class-weighted CrossEntropyLoss for Tasks 1 and 3, and a two-stage cascade architecture for Task 2. Our experiments demonstrate clear improvement over the SVM baseline on Tasks 1 and 3, achieving Macro F1-scores of 0.8417 and 0.7798 respectively. Task 2 remains a hard open problem driven by the extreme scarcity of Violence-labeled samples; our cascade system reaches 0.4022 Macro F1, partially recovering from the total failure of the baseline on the Violence class.

The results confirm that monolingual Spanish Transformer models are effective for sensitive, domain-specific NLP tasks in song lyrics, but that extreme class imbalance at the fine-grained type level requires dedicated augmentation and retrieval strategies beyond standard weighting and oversampling.

Acknowledgments

We dedicate this work to the master's degree scholarship program in Engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia. We want to express our gratitude to the team at the VerbaNex AI Lab[1] for their dedication, collaboration, and ongoing support of our research endeavors.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) to assist with LaTeX formatting and manuscript editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

[1] https://github.com/VerbaNexAI

References

[1] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting Misogyny in Spanish Tweets: An Approach Based on Linguistics Features and Word Embeddings, in: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 57–64.

[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-WS.org, 2026.

[3] 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).

[4] 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.

[5] 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).

[6] M. J. Sosa Borrero, J. E. Serrano, J. C. Martinez-Santos, VerbaNexAI at CheckThat! 2025: Fine-Tuning DeBERTa for Multi-Label Scientific Discourse Detection in Tweets, in: CEUR Workshop Proceedings, 2025.

[7] D. Almanza, J. Martinez Santos, E. Puertas, VerbaNexAI at SemEval-2025 Task 11 Track A: A RoBERTa-Based Approach for the Classification of Emotions in Text, in: Proceedings of SemEval-2025, 2025.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: Proceedings of the 2020 Workshop on Economic Policy and NLP (ECONLP 2020) at ICLR, 2020.

[9] 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. Rush, Transformers: State-of-the-art natural language processing, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020) 38–45.