VerbaNex AI Lab at WomenHelp 2026: BETO-Based Classification of Gender-Based Violence Reports in Spanish

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 WomenHelp 2026 shared task at IberLEF 2026, which focuses on the automatic classification of anonymized gender-based violence (GBV) reports written in Spanish from northern Mexico. We address both subtasks: Subtask 1, which requires classifying the degree of violence on a four-level ordinal scale (mild, medium, high, and severe), and Subtask 2, which requires multi-label classification of the type of violence across seven non-exclusive categories. Our approach progresses from a baseline using TF-IDF representations and XGBoost / MLP models to a Transformer-based system fine-tuned on BETO (`dccuchile/bert-base-spanish-wwm-uncased`), a monolingual Spanish BERT model. To handle severe class imbalance in Subtask 1, we employ a custom Hugging Face `Trainer` with softened class weights and Early Stopping. For Subtask 2, BETO is configured with `problem_type="multi_label_classification"`, which automatically applies BCEWithLogitsLoss for independent per-label predictions. Our final system achieves a Macro F1-score of 0.5552 on Subtask 1 (12th out of 16 teams) and a Micro F1-score of 0.8782 on Subtask 2 (2nd out of 15 teams), demonstrating the effectiveness of monolingual Spanish Transformers for sensitive, domain-specific NLP tasks.

Keywords

gender-based violence, Spanish NLP, BETO, text classification, multi-label classification, class imbalance, IberLEF 2026

1. Introduction

Gender-based violence (GBV) is one of the most serious human rights challenges worldwide, and the automatic detection and classification of GBV narratives has become a growing area of interest in Natural Language Processing (NLP) [1]. The WomenHelp 2026 shared task [2], organized within the IberLEF 2026 evaluation campaign [3], addresses this challenge by providing a corpus of anonymized narratives derived from real GBV reports from northern Mexico, collected during institutional victim care processes.

The task is divided into two subtasks. Subtask 1 requires classifying the *degree* of violence described in each report along a four-level ordinal scale: mild (Class 0), medium (Class 1), high (Class 2), and severe (Class 3). Subtask 2 requires identifying the *type* of violence, where a report may simultaneously belong to one or more of seven non-exclusive categories: economic, physical, patrimonial, psychological, sexual, vicarious, and not applicable (N/A).

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 [4] and emotion classification in Spanish with RoBERTa at SemEval-2025 [5]. This experience with fine-tuning under class imbalance and multi-label settings directly motivated our approach to WomenHelp 2026.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

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)

The primary challenges of this task are: (1) severe class imbalance, with the Severe class comprising only approximately 4% of Subtask 1 training data; (2) domain-specific language of formal institutional and legal records; and (3) the morphological richness of Mexican Spanish. We address these with a pipeline that combines domain-adapted text preprocessing, class imbalance handling, and fine-tuning of BETO [6], a large-scale 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

WomenHelp 2026 provides annotated corpora of GBV reports in Spanish. The reports were collected from institutional victim care processes and manually anonymized to remove personal identifying information while preserving the regional linguistic variation and formal register of official documents.

Subtask 1 (Classification by Degree of Violence) is a four-class single-label classification task. Classes are defined on an ordinal severity scale: mild (0), medium (1), high (2), and severe (3). The evaluation metric is Macro F1-score, which weighs all classes equally regardless of support, making the minority severe class as important as the majority classes.

Subtask 2 (Classification by Type of Violence) is a multi-label classification task with seven non-exclusive labels: L0 (Economic), L1 (Physical), L2 (Patrimonial), L3 (Psychological), L4 (Sexual), L5 (Vicarious), and L6 (Not Applicable). A single report may carry multiple labels. Evaluation metrics are Micro F1-score and Hamming Loss.

2.2. Dataset

The training corpus for Subtask 1 contains approximately 17,000 records; Subtask 2 contains approximately 6,000. Table 1 summarises the dataset statistics for both subtasks and competition phases.

Table 1
WomenHelp 2026 dataset statistics.

Subtask	Split	Records	Labels
ST1	Train	∼17,000	4 classes
ST1	Dev / Test	held-out	—
ST2	Train	∼6,000	7 labels
ST2	Dev / Test	held-out	—

The Subtask 1 distribution is heavily imbalanced: Class 3 (Severe) comprises only approximately 4% of training records. This extreme imbalance is the central challenge of Subtask 1, since the severe class is also the most socially critical to detect correctly.

3. System Description

Our approach evolves across two tiers of increasing complexity, mirroring the progression strategy adopted in other VerbaNex AI Lab submissions [4]: a classical machine learning baseline followed by fine-tuned Transformer models. Figure 1 summarises the full pipeline.

3.1. Text Preprocessing

Given the domain-specific nature of the reports — written in formal administrative Spanish, often in uppercase, and containing legal terminology — we designed a specialized preprocessing pipeline applied to the baseline path:

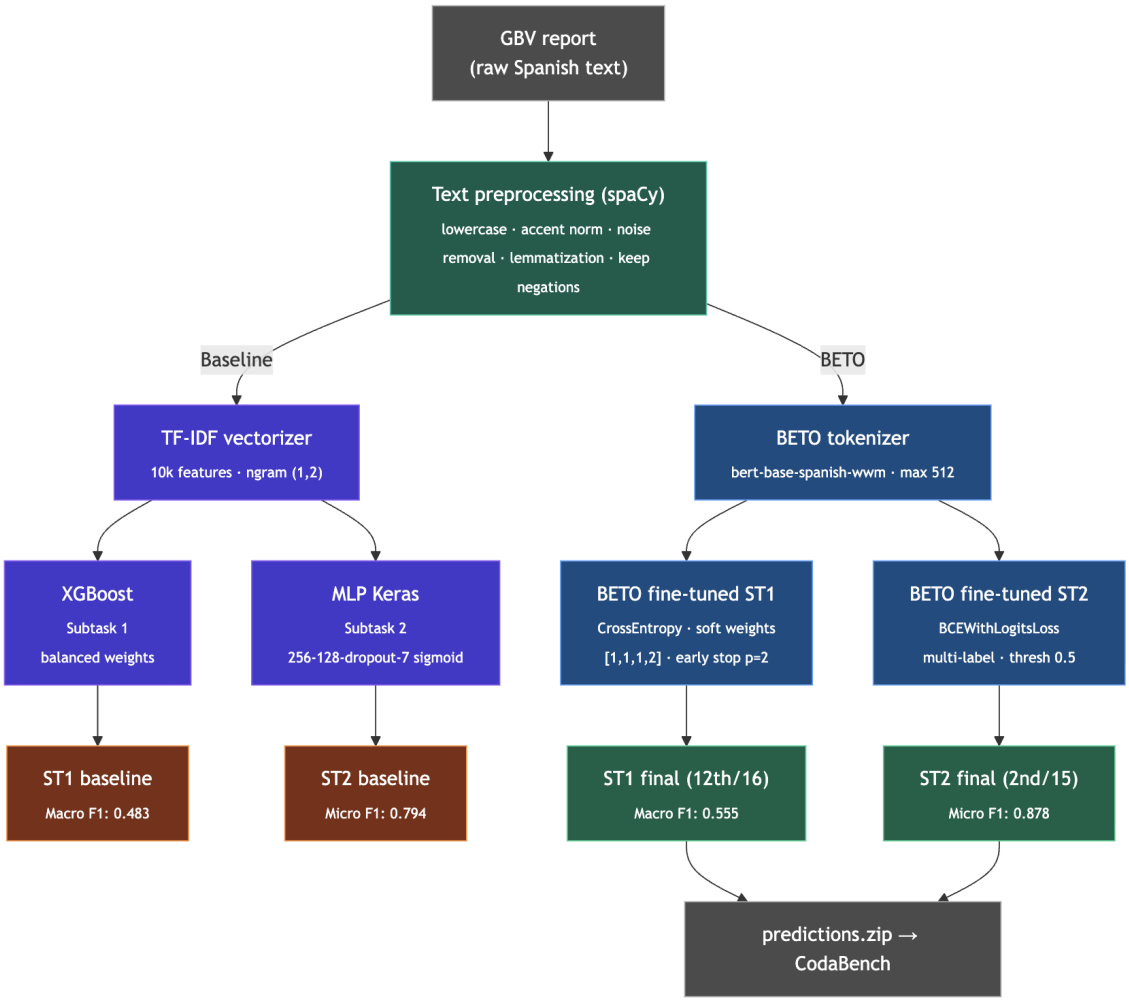

Figure 1: Full system pipeline for both subtasks. The baseline path uses TF-IDF features with XGBoost (ST1) and MLP (ST2). The main system path fine-tunes BETO with task-specific loss functions and class imbalance handling.

1. **Lowercasing**: All text converted to lowercase to reduce vocabulary size.
2. **Accent normalization**: Unicode NFD normalization followed by removal of diacritic marks.
3. **Noise removal**: Punctuation, digits, and special characters removed; multiple spaces collapsed to a single space.
4. **Lemmatization and stopword filtering**: Words are lemmatized using the `es_core_news_md` model from spaCy [7], and standard Spanish stopwords are removed. Semantically critical negation tokens (*no*, *ni*, *sin*, *nunca*) are explicitly preserved, as their presence is highly relevant in descriptions of violence events.

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.

3.2. Baseline: TF-IDF + XGBoost / MLP

We note that the baseline reported here is our own implementation, distinct from the official baseline provided by the shared task organizers. We designed it as an internal lower-bound reference using classical machine learning, in order to quantify the gain obtained from the Transformer-based system on our own preprocessing pipeline and validation split.

Subtask 1 baseline. We trained an XGBoost classifier [8] on TF-IDF features (`max_features=10000`, `ngram_range=(1,2)`). To counteract class imbalance, per-sample

weights were computed using scikit-learn's `compute_sample_weight('balanced')`, which applies the mathematically balanced inverse-frequency weighting. The model was configured with 300 estimators, maximum depth 6, and learning rate 0.1. Importantly, this balanced weighting scheme differs from the softened class weights used in our BETO system (Section 3.3.1); we discuss the implications of this difference in Section 4.3.

Subtask 2 baseline. For the multi-label setting we built a four-layer Keras MLP: Dense(256, ReLU) → Dense(128, ReLU) → Dropout(0.3) → Dense(7, Sigmoid), compiled with binary cross-entropy loss and the Adam optimizer, trained for 15 epochs with batch size 32. A decision threshold of 0.5 was applied to sigmoid outputs.

3.3. Main System: BETO Fine-tuning

Our primary system fine-tunes BETO (`dccuchile/bert-base-spanish-wwm-uncased`) [6], 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 and its ability to capture the morphological richness of the language. Both subtasks share the same base model but use distinct fine-tuning configurations. Figure 2 illustrates the architecture for each subtask.

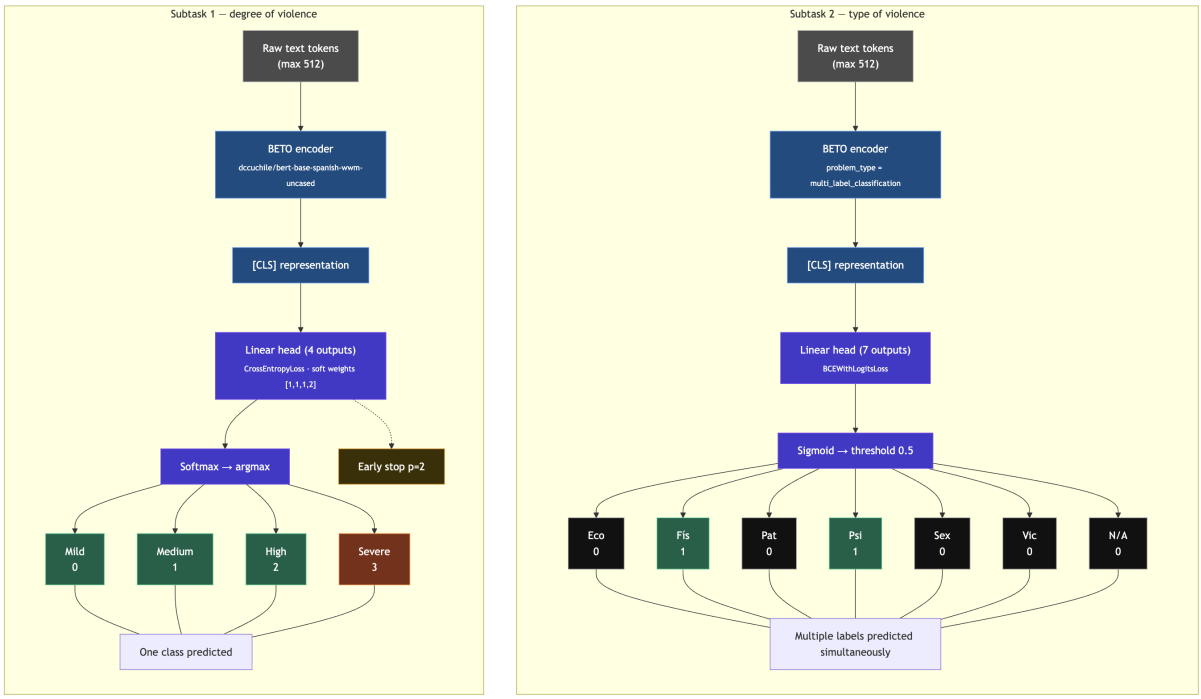

Figure 2: BETO fine-tuning architecture. Left: Subtask 1 uses a 4-class linear head with CrossEntropyLoss and soft class weights [1,1,1,2]. Right: Subtask 2 uses a 7-output linear head with BCEWithLogitsLoss under `multi_label_classification` mode and a sigmoid threshold of 0.5.

3.3.1. Subtask 1: Degree of Violence Classification

BETO is fine-tuned for 4-class sequence classification. We address the class imbalance problem with two complementary strategies.

Softened class weights. Rather than using mathematically balanced weights (which can over-penalize majority classes and destabilize training), we apply manually tuned soft weights $[1.0, 1.0, 1.0, 2.0]$ for the mild, medium, high, and severe classes respectively. These weights are injected into the loss function via a custom `SoftWeightedTrainer` that subclasses Hugging Face `Trainer`:

```
class SoftWeightedTrainer(Trainer):
  def compute_loss(self, model, inputs,
                   return_outputs=False, **kwargs):
    labels = inputs.get("labels")
    outputs = model(**inputs)
    logits = outputs.get("logits")
    loss_fct = nn.CrossEntropyLoss(
        weight=torch.tensor([1.,1.,1.,2.]).to(device))
    loss = loss_fct(
        logits.view(-1, self.model.config.num_labels),
        labels.view(-1))
    return (loss, outputs) if return_outputs else loss
```

Early Stopping. We configured `EarlyStoppingCallback` with patience 2: training stops when the validation Macro F1-score does not improve for two consecutive epochs, preventing overfitting on majority classes. Training hyperparameters: learning rate 2×10^{-5}, batch size 8, up to 6 epochs, weight decay 0.01, max sequence length 512 tokens.

3.3.2. Subtask 2: Multi-label Violence Type Classification

For the multi-label setting, BETO is configured with `problem_type="multi_label_classification"`, which instructs Hugging Face to automatically replace the default CrossEntropyLoss with BCEWithLogitsLoss — the appropriate loss for independent binary predictions across labels [9]. Labels are represented as float vectors of length 7. The `compute_metrics` function applies a sigmoid function to raw logits and a decision threshold of 0.5:

```
def compute_metrics_s2(eval_pred):
    logits, labels = eval_pred
    probs = 1 / (1 + np.exp(-logits))
    preds = (probs >= 0.5).astype(int)
    labels = labels.astype(int)
    return {"f1": f1_score(labels, preds,
                           average="macro")}
```

Training hyperparameters: learning rate 2×10^{-5}, batch size 8, 3 epochs, weight decay 0.01, max sequence length 512 tokens.

4. Experiments and Results

We report results across two evaluation stages: the CodaBench development phase (where we submitted two rounds) and the official testing phase.

4.1. Development Phase

Table 2 shows the scores on the CodaBench development phase leaderboard across our two submission rounds.

The BETO system consistently outperforms the TF-IDF baseline on both subtasks. The most notable improvement is in Subtask 2, where the Micro F1-score increases by 7.8 points and Hamming Loss drops from 0.1123 to 0.0727, confirming that contextual representations from BETO are substantially more effective than TF-IDF bag-of-words for multi-label violence type classification.

4.2. Final Testing Phase

Table 3 presents the official results on the testing phase and the final team ranking.

Our system achieves a strong 2nd place in Subtask 2, with a Micro F1-score of 0.8782 and Hamming Loss of 0.0692 — only 0.0097 Micro F1 points behind the top system. In Subtask 1, the Macro F1 of 0.5552

Table 2
Results on the CodaBench development phase.

Submission	Subtask	Metric	Score
Submission 1 (Baseline)	ST1	Macro F1	0.4830
	ST1	Micro F1	0.5410
	ST2	Micro F1	0.7939
	ST2	Hamming Loss	0.1123
Submission 2 (BETO)	ST1	Macro F1	0.5162
	ST1	Micro F1	0.5477
	ST2	Micro F1	0.8718
	ST2	Hamming Loss	0.0727

Table 3
Official results on the testing phase and final ranking.

Subtask	Macro F1	Micro F1	Hamming Loss	Rank	Teams
ST1	0.5552	0.5715	—	12th	16
ST2	0.8177	0.8782	0.0692	**2nd**	15

places us 12th, consistent with the general difficulty of handling the severe violence minority class: the top-ranking system in Subtask 1 reached only 0.6272.

4.3. Discussion

The divergence in ranking between subtasks reveals an important property of the two tasks. Subtask 2 benefits greatly from BETO's contextual representations: the model can capture co-occurrence signals between violence types within the same narrative — for example, economic and psychological violence frequently co-occur — through attention over the full sequence. This is consistent with the general finding from VerbaNex AI Lab's experience with multi-label classification tasks, where Transformer-based encoders consistently outperform bag-of-words approaches by a wide margin [4, 5].

Subtask 1 is harder because the four ordinal classes are linguistically subtle: the difference between a *high* and *severe* report may depend on specific phrases, legal terminology, or contextual factors that require deeper semantic grounding. The soft class weighting mitigates but does not fully resolve the imbalance problem.

It is worth highlighting that the class imbalance treatment is not identical across our two model families. The baseline applies scikit-learn's mathematically balanced weights (inverse class frequency), which aggressively up-weights the minority Severe class, whereas the BETO system uses manually softened weights $[1, 1, 1, 2]$ that apply a more moderate correction. This difference in weighting regime affects the comparability of the two systems on Subtask 1: the balanced scheme can push the baseline toward higher recall on the Severe class at the cost of precision, while the softened scheme favours overall stability. A fully controlled comparison would require applying the same weighting strategy to both models, which we leave for future work. Future work could also explore ordinal regression formulations or data augmentation strategies for the severe class.

4.4. Error Analysis on Subtask 1

To better characterize the behaviour of our BETO system on the degree-of-violence task, we computed the confusion matrix on the held-out validation split (2,045 reports). Figure 3 shows the matrix normalized by true class, and Table 4 reports per-class precision, recall, and F1.

The error structure is markedly *ordinal*: misclassifications overwhelmingly fall on severity levels adjacent to the true class, rather than scattering across distant classes. This is most visible for the

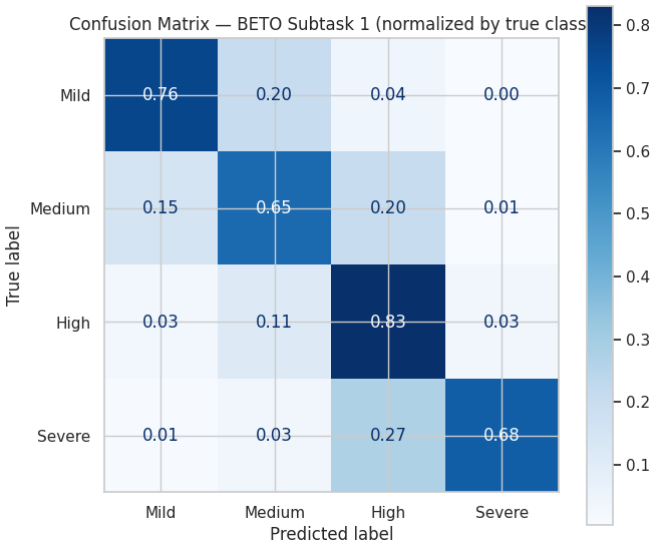

Figure 3: Confusion matrix for the BETO system on the Subtask 1 validation split, normalized by true class. Most confusion concentrates between adjacent severity levels; the Severe class is most often confused with High.

Table 4
Per-class performance of the BETO system on the Subtask 1 validation split.

Class	Precision	Recall	F1	Support
Mild (0)	0.7216	0.7590	0.7398	444
Medium (1)	0.7206	0.6456	0.6810	711
High (2)	0.7828	0.8308	0.8061	798
Severe (3)	0.6702	0.6848	0.6774	92
Macro avg	0.7238	0.7300	0.7261	2045

minority Severe class, where 68.5% of reports are correctly identified but 27.2% are misclassified as High, and only 3.3% and 1.1% fall as far as Medium and Mild respectively. In other words, the model rarely makes catastrophic errors on Severe reports (e.g. predicting Mild); when it errs, it underestimates severity by a single level. A symmetric pattern appears for the Medium class, which is confused with both adjacent levels (20.0% predicted as High, 14.9% as Mild).

Inspecting the misclassified Severe reports reveals a plausible cause: many describe escalating physical aggression and confinement in language that overlaps substantially with High-severity reports, and the decisive cues that distinguish Severe (e.g. explicit threats to life or use of weapons) are sometimes implicit or only inferable from context. This confirms the qualitative observation that the Severe/High boundary is the central difficulty of Subtask 1, and suggests that an ordinal-aware loss or targeted augmentation of borderline Severe examples would be the most promising direction for improvement.

5. Conclusion

We presented the VerbaNex AI Lab system for the WomenHelp 2026 shared task at IberLEF 2026. Our approach fine-tunes BETO for both subtasks: a custom-loss Trainer with softened class weights and Early Stopping for Subtask 1, and a multi-label BCEWithLogitsLoss configuration for Subtask 2. Our experiments demonstrate a clear progression from TF-IDF baselines to the Transformer-based system, yielding 2nd place in Subtask 2 (Micro F1: 0.8782) and 12th place in Subtask 1 (Macro F1: 0.5552).

The results confirm that monolingual Spanish Transformer models are highly effective for sensitive, domain-specific NLP tasks in Spanish, particularly for multi-label type classification. The class imbalance in the degree-of-violence classification task remains an open challenge for the community, where the

socially most critical class (Severe) is also the hardest to detect.

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.

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] N. Martínez-Guevara, G. Bel-Enguix, V. Soto-Mendoza, J. Beltrán, A. Curiel, S. L. Ojeda-Trueba, M. Ramírez-González, H. Gómez-Adorno, Overview of WomenHelp at IberLEF 2026: Classification of Gender-Based Violence Reports to Help Threatened Women in Northern Mexico, Procesamiento del Lenguaje Natural 77 (2026).

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

[4] 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.

[5] 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.

[6] 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.

[7] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear (2017).

[8] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 785–794.

[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.

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