Marcelhr at WomenHelp 2026: Conservative Ensembles and Ordinal Routing for Gender-Based Violence Classification

Marcel Herrera Rendón[1,*]

[1]Master's Program in Data Science, Universidad de Sonora, Hermosillo, Sonora, Mexico

Abstract

Our main contribution to WomenHelp 2026, an IberLEF shared task on anonymized gender-based violence reports from northern Mexico, is a conservative model-selection strategy under distribution shift, rather than a new architecture. The task has two subtasks: ordinal severity classification with four ordered labels and multilabel violence-type classification with seven binary labels. We frame it as empirical model selection, after observing that several candidates that looked strong on the development set failed to generalize to the external (official held-out) test set. Prompt-based and LoRA-based language models were not competitive, whereas Spanish transformer encoders and their ensembles were stronger and more controllable. Our final system pairs a conservative severity router—which anchors on a previously validated ten-model ensemble and only adjusts three selected ordinal transitions—with a fixed, externally validated BETO multilabel classifier. It reached 0.60798 severity macro-F1 (3rd of 16 teams) and 0.87016 violence-type micro-F1 (6th of 15 teams), suggesting that controlled routing around externally validated anchors can be safer than replacing a model with a stronger-looking local candidate.

Keywords

gender-based violence detection, Spanish natural language processing, ordinal text classification, multi-label classification, severity classification, conservative model ensembles, risk-aware model selection, robustness to distribution shift, Spanish transformer encoders (BETO)

1. Introduction

WomenHelp 2026 [1], organized as part of IberLEF 2026 [2], focuses on automatic analysis of textual reports related to violence against women. The practical challenge is not limited to detecting explicit violent vocabulary. A useful system must distinguish severity levels and identify different types of violence in narratives that may contain institutional language, indirect descriptions, repeated events, threats, consequences, and multiple overlapping forms of abuse.

The shared task consists of two subtasks. Subtask 1 predicts the severity of a report using four ordered labels: Mild, Medium, High, and Severe. Although the official metric evaluates this as a standard multiclass problem, the label semantics are ordinal. Subtask 2 predicts seven binary violence-type labels as a multilabel problem. We used macro-F1 as the main model-selection criterion because it is robust to the strong class imbalance in the data; we note, however, that the official violence-type leaderboard was ultimately ranked by micro-F1.

Our development process was driven by three observations. First, prompt-only generative systems were not reliable enough for the required structured predictions. Second, Subtask 2 became relatively stable with discriminative models, while Subtask 1 remained sensitive to class imbalance and distribution shift. Third, several severity candidates that looked strong on the development set failed to generalize externally. Consequently, the final system deliberately prioritized stability over aggressive local optimization.

The main contribution of our submission is therefore not a single new architecture, but a conservative model-selection strategy: retain externally validated components, use ordinal evidence only where it is

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
mariron42@gmail.com (M. H. Rendón)
0009-0006-9702-8852 (M. H. Rendón)

supported, and avoid replacing a strong baseline with a locally overfit candidate.

2. Task and Data

The corpus contains anonymized narratives derived from real reports of gender-based violence in northern Mexico. The official evaluation reports macro-F1, micro-F1, and additional metrics; severity (Subtask 1) was ranked by macro-F1, while violence types (Subtask 2) were ranked by micro-F1, with Hamming loss as a tie-breaker. Subtask 1 additionally reports micro-F1; Subtask 2 additionally reports macro-F1, weighted-F1, and Hamming loss.

Terminology. The task uses three partitions: training, development, and test. Throughout, the *development set* is the released development partition (with gold labels), which the organizers used for the development-phase leaderboard and which we used for local model selection. The *external test set* is the held-out test partition, whose labels were never released to participants and which was scored only through the Codabench submission system in the final phase. Accordingly, an *external submission* is a prediction file scored on that test partition, an *external result* is its returned score, and an *externally validated component* is a model whose score we confirmed on the external test set rather than only on the development set. Every development-set metric we report is locally reproducible from released data, whereas all external results are official platform-returned scores that cannot be recomputed locally, because the test labels were never released; the final ranking is publicly documented on the competition results page [3].

Subtask 1 has four severity labels: Mild, Medium, High, and Severe. During modeling, we treated these labels as an ordinal scale, even when using standard multiclass losses. Subtask 2 has seven binary labels: Economic, Physical, Property-related (Patrimonial in the official label set), Psychological, Sexual, Vicarious, and N/A. The violence-type labels were produced by three human annotators: a label is positive when at least two of the three agree (a mean annotation above 0.66), and the organizers additionally released the per-document mean annotations as auxiliary soft labels, which motivated our later soft-label experiments.

The training and development sets were released first, as a single labeled bundle, whereas the held-out test set was released later, without labels, only for the final evaluation phase. The released files contain 13630 training and 3405 development reports for Subtask 1, and 4845 training and 1212 development reports for Subtask 2. The held-out test partition, on which the final leaderboard was computed, contains 4219 reports for Subtask 1 and 3620 for Subtask 2. The two subtasks did not share exactly the same rows, and direct alignment between Subtask 1 and Subtask 2 reports was limited. This affected feature design: although violence types are semantically informative for severity, we avoided relying on direct Subtask-2-to-Subtask-1 predictions in the final submission.

The public split is markedly imbalanced. For Subtask 1, the official class distribution is approximately 22% Mild, 35% Medium, 39% High, and 4% Severe; concretely, Severe represented only 616 of 13630 training reports and 770 of 17035 train-plus-development reports. For Subtask 2, rare labels included N/A, Vicarious, and Sexual. These imbalances shaped our use of class weighting, focal loss [4], targeted augmentation, and macro-F1-based selection.

3. Experimental Development

3.1. Prompt-Based and LoRA-Based Baselines

We first tested instruction-tuned language models with direct prompting. This was a reasonable starting point because the inputs are narrative and require semantic interpretation. It was also motivated by the broad interest in large language models as general-purpose problem solvers in 2026. Our initial hypothesis was not only that an LLM could classify the reports, but that it could return short, structured evidence for each severity and violence-type decision, making the output easier for a human professional

to inspect. In practice, this did not work well enough for the competition setting: the models were weaker classifiers than discriminative baselines, often overestimated severity or produced overly broad labels, and frequently failed the strict structured-output contract.

The main prompt baseline used the Gemma 4 E2B instruction model, implemented through the `unsloth/gemma-4-E2B-it` checkpoint. It used deterministic decoding (`temperature=0.0`, `top_p=0.9`), `max_new_tokens=96`, `min_new_tokens=16`, `max_seq_length=2048`, and a `High` fallback for missing severity predictions. This baseline produced 0.3085 macro-F1 for Subtask 1 and 0.5590 macro-F1 for Subtask 2. A simple classical hashing baseline already improved Subtask 1 to 0.4371 macro-F1.

We also explored LoRA fine-tuning of the same Gemma 4 E2B instruction model. The LoRA configuration used `max_seq_length=1024`, batch size 1 with gradient accumulation 8, learning rate `1e-5`, `stable_adamw`, 400 maximum steps, fp32 training, seed 3407, rank 16, alpha 32, and target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The seed was a fixed reproducibility seed carried across experiments, not a tuned hyperparameter. That path was ultimately discarded. Several runs had training instability or output-format failures. A longer adapter evaluation completed, but the model collapsed to empty or unparsable outputs, producing a valid-JSON rate of 0.0. This made the generative route too risky for the shared task, where a malformed output is equivalent to a failed prediction.

This phase changed our modeling direction. Instead of asking a generative model to produce structured answers, we moved to discriminative classifiers that produced scores, probabilities, and auditable CSV files. Nevertheless, we consider evidence-generating LLMs a promising future research direction if the output format and calibration problems can be solved.

3.2. Classical TF-IDF Models

Classical models were stronger than expected, especially for Subtask 2. We evaluated word TF-IDF, character TF-IDF, word-character combinations, and character word-boundary n-grams with classifiers such as Logistic Regression, Linear SVM, SGDClassifier, Complement Naive Bayes, KNN, and ExtraTrees.

For Subtask 1, the best early classical severity system used word-character TF-IDF, ordinal modeling, and auxiliary violence-type signals. It reached 0.4974 development macro-F1. This was a useful improvement over the initial baselines, but it remained below later encoder and ensemble systems.

For Subtask 2, a character SVM became an important reference, reaching 0.7642 macro-F1, 0.8449 micro-F1, and 0.0903 Hamming loss on development data. These models remained useful even after transformer encoders were introduced, because they captured lexical patterns that were complementary to contextual encoders.

3.3. Transformer Encoders for Severity

We trained several Spanish and multilingual encoders, including BETO [5], BETO uncased, BERTin [6], Electricidad, XLM-R base, and XLM-R large [7]. On the Yuca cluster with AMD MI210 GPUs and ROCm, stable training required practical changes: eager attention, fp32 precision, `adamw_torch`, gradient clipping, and careful handling of model caches. Some candidate backbones, such as RobERTuito in our environment, failed with ROCm memory access errors before completing training.

In this paper, Electricidad refers to the `mrm8488/electricidad-base-discriminator` checkpoint, a Spanish ELECTRA-style discriminator encoder available on Hugging Face [8, 9]. It is not a model introduced by us and was not used generatively; we fine-tuned it as a supervised severity encoder. Its main role was to add ensemble diversity, especially in focal-loss variants, rather than to outperform BETO as a standalone model.

We selected BETO uncased as the primary discriminator for three reasons: it is pre-trained on large Spanish corpora and therefore matches the language of the reports; the narratives are frequently transcribed in upper case or with noisy, inconsistent casing, so an uncased model is more robust than a cased one; and across our runs it produced the strongest and most stable individual macro-F1.

Adapting these encoders to the task required several task-specific changes rather than off-the-shelf use. For severity, we added a linear classification head over the pooled representation and trained it with class-weighted cross-entropy and focal loss [4] to counter the Severe minority, treating the four labels as an ordinal scale at decision time. For violence types, we replaced the softmax head with seven independent sigmoid outputs and tuned a decision threshold per label (or a single fixed threshold in the final submission), because the labels co-occur and are not mutually exclusive. These adaptations, together with the ROCm-related training changes above, were what made the encoders usable and competitive on this corpus.

The early encoder experiments showed that BETO-based models were stronger than initial XLM-R base runs. In the larger batch, BETO uncased became the strongest individual encoder family. Representative Subtask 1 development results are shown in Table 1.

Table 1
Representative individual encoder results for Subtask 1.

System	Macro-F1	Accuracy
BETO uncased + focal loss	0.5429	0.5712
BETO uncased + cross-entropy	0.5360	0.5765
Electricidad + focal loss	0.5216	0.5586
BERTin + cross-entropy	0.5176	0.5574

The conclusion was that encoders were necessary but not sufficient. No individual encoder solved the severity task, especially the High-Severe boundary.

3.4. Mega-Ensembles for Severity

The best early Subtask 1 system was not a single encoder but a ten-model severity ensemble. It combined three BETO uncased variants, three Electricidad variants, three classical severity ensemble components, and one BERTin variant.

This ensemble reached 0.5589 development macro-F1 and became the main severity anchor. We refer to it as the Top10 severity anchor throughout the rest of the paper. A simpler three-model mean ensemble reached 0.5550, while the best individual encoder remained at 0.5429. This showed that model diversity mattered more than selecting a single best backbone. Exact experiment identifiers are preserved in the repository metadata rather than used as primary names in the paper.

However, the ensemble also introduced a reproducibility issue. Some later candidates depended on reconstructed classical probability sources that drifted under the current library stack. We therefore separated high-scoring search candidates from faithfully materializable candidates. This distinction became important in the final submission.

3.5. Ordinal Severity Modeling

Because Subtask 1 is semantically ordinal, we tested cumulative-score mixtures and threshold-based ordinal predictions. One reproducible-only ordinal search excluded known drifting sources and combined ten materializable predictions with equal weights. The best reproducible ordinal ensemble achieved 0.580536 development macro-F1, 0.608517 micro-F1, 0.473684 Severe F1, and 0.409091 Severe recall, compared with 0.558879 macro-F1 for the Top10 severity anchor.

A paired bootstrap over the development set produced a positive delta in 99.4% of samples. Locally, the ordinal candidate looked strong. But it changed 636 of 4219 official-test severity predictions relative to the previous best external submission, and it shifted many Mild cases toward Medium or High. This made it risky.

We then introduced drift-penalized search. This variant reduced distribution drift and looked more balanced in slice validation. Nevertheless, its external result was worse than the previous best submission: 0.601431 macro-F1 versus 0.606777. This was one of the most important negative results of the project.

Even with bootstrap, slice validation, and drift penalties, the severity development set remained optimistic. For the final attempt, we stopped trying to replace the full severity system and switched to a limited routing strategy.

3.6. Targeted Data Augmentation

Data augmentation was tested because Severe was underrepresented and because severity boundaries depended on report style, not only isolated keywords. We found that blind augmentation was risky, while targeted augmentation could be useful as a diversity source.

The clearest positive result came from report-style augmentation for a BETO uncased focal-loss severity encoder. The BETO focal baseline reached 0.548114 macro-F1, 0.583847 micro-F1, and 0.417808 Severe F1. BETO with report-style augmentation improved these scores to 0.555198, 0.587372, and 0.433460, respectively.

The report-style augmentation model was not only stronger individually; it appeared in half of the top 20 diverse ordinal ensembles. The best ensemble containing it reached 0.575680 development macro-F1, compared with 0.573706 for the best top-ranked ensemble without it. We also tried Mild-oriented augmentation and Severe synonym augmentation. These variants helped diversity, but the report-style augmentation was the cleanest signal.

3.7. Subtask 2 Improvements and Non-Promoted Variants

Subtask 2 was strong enough that we used a higher bar for replacing the current component. The externally validated system used BETO uncased with balanced weighting and a fixed threshold of 0.45. It reached 0.808022 external macro-F1, 0.870165 micro-F1, and 0.075770 Hamming loss.

We still tested whether out-of-fold thresholding, classical OVR diversity, soft labels, or asymmetric loss [10] could improve the multilabel component. The OOF BETO experiment improved train-fold OOF scores, but its thresholds did not transfer to development. A BETO+OVR fusion improved development macro-F1 by only 0.001943, which was too small to justify a limited submission. The soft-label model improved micro-F1, weighted-F1, and Hamming loss, but it reduced macro-F1. Because our internal replacement criterion prioritized macro-F1 and the soft-label variant reduced it, we did not promote it; under the final official micro-F1 ranking for Subtask 2, this is better read as a trade-off than as a definitive negative result. We also note that these soft-label runs were carried out after the final submission, so they could not have informed it.

4. Final System

4.1. Subtask 1: Conservative Severity Router

The final severity submission used the previous best external Subtask 1 output as an anchor. We then applied a post-hoc conservative router: a rule-based decision layer that could override the anchor only when an alternative foldbag candidate crossed an ordinal threshold and the proposed change belonged to a small set of allowed transitions. This made the router different from a full ensemble replacement; most predictions remained exactly equal to the anchor.

The candidate foldbag consisted of five folds of BETO with report-style augmentation, five folds of Electricidad with focal loss, equal family weights of 0.5 and 0.5, and ordinal thresholds [0.6, 0.55, 0.55]. The family weights and ordinal thresholds were selected empirically from local development searches, but the final choice also required materialization checks and conservative external-risk review. The thresholds correspond to the three cumulative severity boundaries and were not allowed to trigger arbitrary non-ordinal jumps.

The router allowed only three transitions: Medium to Mild, Mild to Medium, and High to Severe. The High-to-Severe transition used an additional severe-evidence condition based on high-risk lexical cues, including threats, weapons, sexual violence, and severe consequences. Thus, the model could not freely

promote a report to the highest severity level without lexical high-risk cues in the text. The router did not replace the full severity prediction; it only changed cases in selected transition regions. Figure 1 summarizes the decision flow, and its actual impact on the official test set is reported together with the final results.

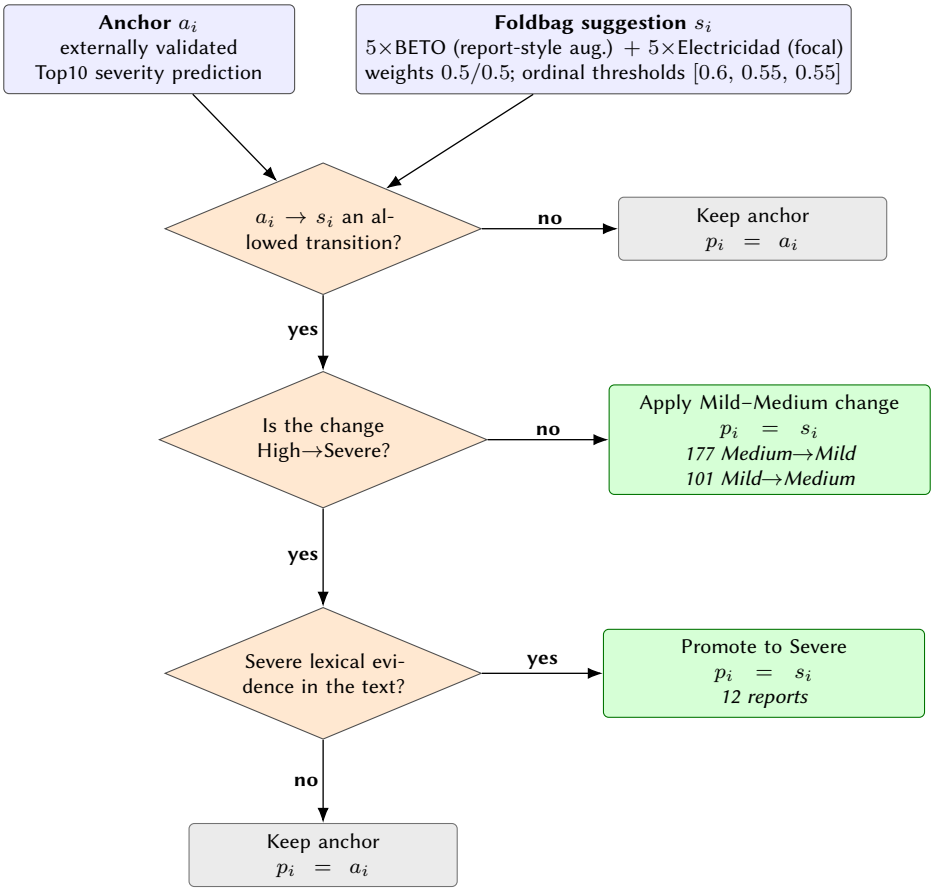

Figure 1: Decision flow of the conservative severity router, annotated with the number of official-test reports (of 4219) that follow each branch. For every report the router compares the externally validated anchor a_i with the foldbag suggestion s_i and overrides the anchor only for three allowed ordinal transitions—Medium→Mild, Mild→Medium, and High→Severe—so the dominant Medium–High confusion is left untouched by design. High→Severe additionally requires severe lexical evidence in the text (threats, weapons, sexual violence, severe consequences). The router changed 290 predictions in total; the remaining 3929 reports kept the anchor unchanged.

4.2. Subtask 2: Fixed BETO Multilabel Component

The final Subtask 2 output was copied unchanged from the best externally validated BETO-based component. This was a deliberate decision. Later experiments suggested possible gains in Hamming loss or micro-F1, but none provided enough macro-F1 evidence to justify replacing a strong external component.

5. Results

The final external results are shown in Table 2. On the official test set, the router changed 290 of 4219 severity predictions: 177 Medium-to-Mild changes, 101 Mild-to-Medium changes, and 12 High-to-Severe changes. The resulting distribution changed from 829/1964/1310/116 for Mild/Medium/High/Severe in the anchor to 905/1888/1298/128 in the final system. This behavior matched the intended risk profile: the system adjusted the Mild-Medium boundary and added a small number of Severe predictions without

globally replacing the severity model. Compared with our previous best submission, the final system slightly improved Subtask 1 while preserving Subtask 2 exactly (Table 3).

Table 2
External results of the final submission.

Metric	Score
S1 macro-F1	0.607976891066
S1 micro-F1	0.604645650628
S2 macro-F1	0.808022474168
S2 micro-F1	0.870164998648
S2 weighted-F1	0.876879684108
S2 Hamming loss	0.075769534333

Table 3
Comparison against the previous best submission.

Metric	Previous	Final	Difference
S1 macro-F1	0.606777	0.607977	+0.001200
S1 micro-F1	0.602275	0.604646	+0.002370
S2 macro-F1	0.808022	0.808022	0.000000
S2 micro-F1	0.870165	0.870165	0.000000
S2 weighted-F1	0.876880	0.876880	0.000000
S2 Hamming loss	0.075770	0.075770	0.000000

The gain was small, but it was external and did not damage the stable multilabel component.

5.1. Official Ranking

On the official leaderboard [3], our team (`marcelhr`) placed 3rd of 16 teams in Subtask 1 (severity, ranked by macro-F1) and 6th of 15 teams in Subtask 2 (violence types, ranked by micro-F1 with Hamming loss as a tie-breaker). In Subtask 1 we finished 0.0192 macro-F1 below the best system and only 0.0022 above the fourth-placed team, so the conservative strategy was enough for a top-three finish in a tightly packed leaderboard. Notably, the teams that led one subtask did not lead the other: the Subtask 1 winner placed 8th in Subtask 2, and the Subtask 2 winner placed 6th in Subtask 1, which is consistent with our decision to treat the two components separately. Tables 4 and 5 list the top teams per subtask.

Table 4
Subtask 1 leaderboard (top teams, ranked by macro-F1); our team in bold.

Rank	Team	Macro-F1
1	angellabco	0.6272
2	franrodrigo	0.6199
3	**marcelhr (ours)**	**0.6080**
4	arjun108	0.6058
5	omargarcia	0.5936
6	jorgegomezn	0.5915

5.2. Per-Class and Per-Label Analysis

To complement the official scores, we report development-set diagnostics for the two main components. Table 6 and Table 7 show the confusion matrix and per-class results of the Top10 severity anchor. Errors concentrate on the Medium–High boundary (702 confusions across both directions) and on the

Table 5
Subtask 2 leaderboard (top teams, ranked by micro-F1, Hamming loss as tie-breaker); our team in bold.

Rank	Team	Micro-F1	Hamming loss
1	jorgegomezn	0.8879	0.0631
2	verbanexai	0.8782	0.0692
3	phonghoang	0.8740	0.0711
4	aerojash	0.8735	0.0721
5	pln_uam	0.8715	0.0749
6	**marcelhr (ours)**	**0.8702**	**0.0758**

rare Severe class, whose recall stays at 0.41 despite reasonable precision. These diagnostics motivated the conservative router, although its final rule only adjusts the Mild–Medium boundary and a small High-to-Severe subset rather than the dominant Medium–High confusion. Table 8 shows per-label results for the BETO multilabel component at the submitted fixed 0.45 threshold. Frequent labels such as Physical and Psychological exceed 0.91 F1, whereas the rare Vicarious label remains the hardest at 0.59 F1, motivating the label-specific specialists we propose as future work.

Table 6
Subtask 1 confusion matrix on the development set (rows are gold labels, columns are predictions) for the Top10 severity anchor.

Gold ↓ / Pred →	Mild	Medium	High	Severe
Mild	434	286	19	0
Medium	213	595	372	3
High	68	330	890	41
Severe	5	13	73	63

Table 7
Subtask 1 per-class results on the development set (Top10 severity anchor, macro-F1 = 0.5589).

Class	Precision	Recall	F1	Support
Mild	0.6028	0.5873	0.5949	739
Medium	0.4861	0.5030	0.4944	1183
High	0.6573	0.6697	0.6634	1329
Severe	0.5888	0.4091	0.4828	154
Macro avg	–	–	0.5589	3405

Table 8
Subtask 2 per-label results on the development set (BETO uncased multilabel component, fixed 0.45 threshold; macro-F1 = 0.8198, micro-F1 = 0.8685).

Label	Precision	Recall	F1	Support
Economic	0.7391	0.9057	0.8140	244
Physical	0.9258	0.9318	0.9288	616
Property-related	0.6677	0.8474	0.7469	249
Psychological	0.8975	0.9313	0.9141	931
Sexual	0.8771	0.9128	0.8946	172
Vicarious	0.5556	0.6303	0.5906	119
N/A	0.8947	0.8095	0.8500	42
Macro avg	0.7939	0.8527	0.8198	2373

6. Discussion

The main lesson from the project is that local validation was not enough for Subtask 1. Ordinal ensembles, drift penalties, and targeted augmentation all produced plausible local improvements, but some failed externally. The final routing strategy worked because it reduced the decision surface: instead of asking a new model to replace the anchor, it allowed a diverse foldbag to act only on selected transitions.

Subtask 2 behaved differently. The best BETO component was robust enough that most later improvements were marginal or metric-specific. We therefore treated it as a fixed asset. This also reduced risk in the final Hail Mary submission, because any improvement or degradation would come almost entirely from Subtask 1.

The results also support treating severity as ordinal, but not necessarily through a full ordinal replacement model. In our setting, ordinal scores were most useful for controlled routing and error analysis, while aggressive ordinal ensembles overfit the development set. Future work could test formal ordinal neural models such as CORAL [11] or CORN [12], and multilabel dependency models such as classifier chains [13].

7. Limitations and Future Work

The main limitation remains the Medium-High-Severe boundary. Severe is rare and semantically heterogeneous. Some severe cases require understanding threats, consequences, weapons, sexual violence, repeated escalation, or risk to third parties; other high-intensity texts contain strong words but should not necessarily be promoted to Severe.

Future work should prioritize fold-based leave-cluster-out validation for severity, true out-of-fold meta-selection over candidate predictions, formal ordinal neural models, conservative High-vs-Severe specialists with hard negatives, label-specific specialists for rare Subtask 2 labels such as Vicarious and N/A, and better use of official soft labels without sacrificing macro-F1.

8. Declaration on Generative AI

Generative AI assistance was used during system development to brainstorm modeling ideas, organize technical documentation, inspect experiment logs, and draft parts of the project reports and manuscript. The final predictions were produced by trained models and reproducible scripts in the repository. The author manually reviewed the modeling decisions, reported metrics, and manuscript content.

9. Conclusions

Our final WomenHelp 2026 system combined a conservative severity router with a fixed BETO multilabel classifier. The project began with prompt-based and LoRA-based experiments, moved through classical and encoder models, and ultimately converged on a risk-aware ensemble strategy. The final submission improved external Subtask 1 macro-F1 from 0.60678 to 0.60798 while keeping Subtask 2 fixed at 0.80802 macro-F1, and placed 3rd of 16 teams in severity and 6th of 15 in violence types. The improvement was modest, but it was consistent with our central design choice: in a setting with limited submissions and optimistic development validation, controlled routing around externally validated anchors can be safer than replacing the full model with a stronger-looking local candidate.

Acknowledgments

The author acknowledges the financial support provided through the graduate scholarship awarded by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), CVU No. 2138227. This support was fundamental for the completion of the postgraduate studies related to this research.

The author also acknowledges the use of the high-performance computing infrastructure of the High Performance Computing Area at Universidad de Sonora (ACARUS/Yuca) for running model training, inference, and validation jobs.

References

[1] 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).

[2] 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.

[3] WomenHelp 2026 Organizers, WomenHelp 2026 Shared Task: Final Results, https://www.codabench.org/competitions/14614/, 2026. URL: https://www.codabench.org/competitions/14614/, codabench competition 14614, Final Results page. Accessed 2026-06-20.

[4] 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, 2017, pp. 2980–2988. URL: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. URL: https://arxiv.org/abs/2308.02976. arXiv:2308.02976.

[6] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. González de Prado Salas, M. Romero, M. Grandury, BERTIN: Efficient pre-training of a spanish language model using perplexity sampling, 2022. URL: https://arxiv.org/abs/2207.06814. arXiv:2207.06814.

[7] 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, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[8] M. Romero, Spanish electra by manuel romero, https://huggingface.co/mrm8488/electricidad-base-discriminator/, 2020. URL: https://huggingface.co/mrm8488/electricidad-base-discriminator/.

[9] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[10] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91. URL: https://openaccess.thecvf.com/content/ICCV2021/html/Ridnik_Asymmetric_Loss_for_Multi-Label_Classification_ICCV_2021_paper.html.

[11] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recognition Letters 140 (2020) 325–331. URL: https://arxiv.org/abs/1901.07884. doi:10.1016/j.patrec.2020.11.008.

[12] X. Shi, W. Cao, S. Raschka, Deep neural networks for rank-consistent ordinal regression based on conditional probabilities, 2021. URL: https://arxiv.org/abs/2111.08851. arXiv:2111.08851.

[13] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2009, pp. 254–269. URL: https://doi.org/10.1007/978-3-642-04174-7_17. doi:10.1007/978-3-642-04174-7_17.