Manojkumar00011 at WomenHelp 2026: XLM-RoBERTa with Multi-Sample Dropout and Per-Label Threshold Optimization for Gender-Based Violence Report Classification

Manoj Kumar[1,*], Prabavathy Balasundaram[2] and Deeraj Kumar[1]

[1]Department of Computer Science and Engineering (AI & ML), Easwari Engineering College, Chennai, India
[2]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

This paper presents our system submitted to the WomenHelp 2026 shared task at IberLEF 2026, which focuses on the automatic classification of Spanish-language gender-based violence (GBV) reports from Northern Mexico. We participated in both subtasks: Subtask 1, multi-class severity classification (Mild, Medium, High, Severe), and Subtask 2, multi-label identification of violence types (Economic, Physical, Patrimonial, Psychological, Sexual, Vicarious, N/A). Our approach fine-tunes XLM-RoBERTa-base [1] with attention-masked mean pooling and multi-sample dropout [2]. For Subtask 1, we employ Focal Loss [3] with label smoothing to address severe class imbalance. For Subtask 2, we train on expert-annotated soft labels with weighted Binary Cross-Entropy and R-Drop [4] regularization, followed by per-label threshold optimization that yields an absolute Macro F1 improvement of $+0.017$ on the development set. Our system achieves a Macro F1-score of 0.5565 (11^{th} of 16 teams) on Subtask 1 and a Micro F1-score of 0.8584 with a Hamming Loss of 0.0818 (12^{th} of 15 teams) on Subtask 2.

Keywords

gender-based violence classification, multilingual NLP, XLM-RoBERTa, multi-label classification, focal loss, threshold optimization, soft labels

1. Introduction

Violence against women is a global social emergency whose manifestations span physical, psychological, sexual, economic, and other dimensions. In Mexico, institutional support centers produce detailed written records of gender-based violence (GBV) cases as part of victim-care and legal proceedings. These narratives, if automatically analyzed, could help social workers prioritize cases, identify patterns, and support overloaded legal systems.

The WomenHelp 2026 shared task [5] at IberLEF 2026 [6] presents participants with a corpus of anonymized case reports written in informal, domain-specific Mexican Spanish, annotated by expert social workers. The task asks participants to classify these reports by severity of violence (Subtask 1) and by type of violence (Subtask 2).

Automatic classification of such texts poses several challenges. First, the language is domain-specific and regionally varied, containing colloquialisms characteristic of Northern Mexico. Second, class distributions are highly imbalanced: the Severe class represents only approximately 5% of Subtask 1 samples. Third, Subtask 2 is a multi-label problem with co-occurring violence types, where annotator disagreement is significant enough to warrant soft-label annotations.

Our system (team *manojkumar00011*) addresses these challenges with a unified transformer-based architecture built on XLM-RoBERTa-base, a multilingual model that provides strong Spanish language representations out of the box. Task-specific adaptations include Focal Loss for Subtask 1 imbalance, soft-label training and per-label threshold optimization for Subtask 2, and multi-sample dropout and R-Drop regularization applied throughout.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
manojkumar.ksamy@gmail.com (M. Kumar); prabavathyb@ssn.edu.in (P. Balasundaram); deerajkumar05238@gmail.com (D. Kumar)

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the tasks and datasets. Section 4 presents the methodology, covering data preprocessing, the model architecture, the training procedure, and a worked example tracing real reports through the pipeline. Section 5 describes the experimental setup. Section 6 reports official results. Section 7 discusses findings. Section 8 concludes, and Section 9 states the limitations of the work.

2. Related Work

Transformer-based Text Classification. The introduction of BERT [7] transformed natural language processing by enabling deep bidirectional contextual representations through masked language modeling on large unlabeled corpora. RoBERTa [8] refined this recipe by removing the next-sentence prediction objective, increasing batch sizes, and training on more data, yielding consistently stronger downstream results. For multilingual settings, XLM-RoBERTa [1] scales RoBERTa to 100 languages by pretraining on 2.5 TB of filtered CommonCrawl text. This makes it particularly suitable for our task, where reports are written in domain-specific Mexican Spanish that may be underrepresented in Spanish-only pretrained models.

Violence and Harmful Content Detection in NLP. Automated detection of violence and harmful content in text has received increasing attention in the Spanish NLP community. Shared tasks at IberEval and IberLEF have featured dedicated tracks on aggressive language identification, misogyny detection, and gender-based violence classification [5], establishing benchmarks for evaluating systems on Spanish social media and institutional texts. A common challenge across these tasks is severe class imbalance and domain-specific vocabulary. Our work addresses both by applying Focal Loss to counteract the scarcity of severe-class instances, and by leveraging XLM-RoBERTa's broad multilingual pretraining for robust Spanish representations.

Multi-Label Text Classification. Multi-label classification, where each instance may simultaneously belong to multiple categories, presents distinct challenges compared to single-label settings [9]. Key issues include label correlation, per-label class imbalance, and the choice of decision threshold. The Binary Relevance paradigm treats each label as an independent binary problem, enabling the use of per-label weights and thresholds. Our Subtask 2 system follows this paradigm and extends it with soft-label supervision derived from annotator agreement scores, which communicates label uncertainty to the model during training and has been shown to improve calibration in multi-annotator settings.

3. Task and Data Description

3.1. Subtask 1: Classification by Degree of Violence

Given a GBV narrative, systems must assign one of four severity classes: Mild (0), Medium (1), High (2), or Severe (3). The dataset comprises 13,630 training, 3,405 development, and 4,219 test instances (21,254 in total). Class distribution is imbalanced: Mild ≈22%, Medium ≈35%, High ≈39%, and Severe ≈5%. The primary evaluation metric is Macro F1-score, which equally weights all four classes regardless of frequency.

3.2. Subtask 2: Types of Violence

Each report may simultaneously belong to one or more of seven categories: Economic (L0), Physical (L1), Patrimonial (L2), Psychological (L3), Sexual (L4), Vicarious (L5), and N/A (L6). The dataset contains 4,845 training, 1,212 development, and 3,620 test instances (9,677 in total). Labels were set by inter-annotator agreement among three expert social workers; a label is assigned 1 when the mean annotation score is at least 0.67 (i.e., at least two of three annotators agreed). A soft-label variant containing the raw mean

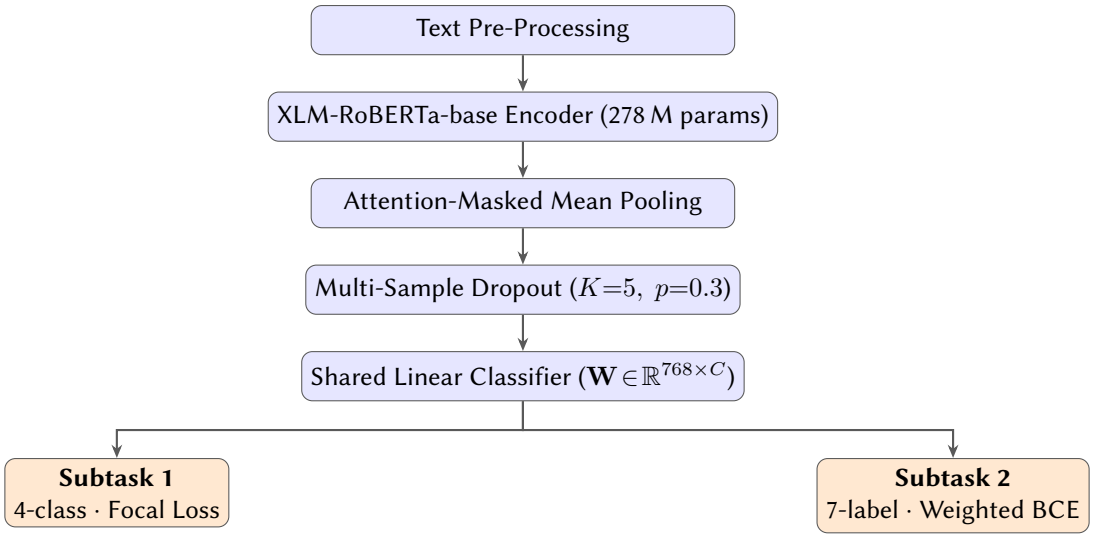

Figure 1: End-to-end system architecture. The input is a raw gender-based violence report written in Mexican Spanish. The preprocessing, encoder, pooling, and multi-sample dropout layers are shared; the loss function and output dimensionality C differ per subtask.

annotation scores (values in $\{0, 0.33, 0.67, 1\}$) is provided as auxiliary data. Evaluation uses Micro F1-score and Hamming Loss (the fraction of label–instance pairs that are predicted incorrectly, for which lower is better).

3.3. Dataset Statistics

Table 1 reports the distribution of positive labels in the Subtask 2 training set. Psychological violence (L3) is the most prevalent category, present in 79.5% of all training instances, while N/A (L6) is the rarest at 3.4%. Physical (L1) and Psychological (L3) violence frequently co-occur, reflecting the compound nature of most reported cases. This extreme label imbalance across categories directly motivates the per-label positive class weights applied in the loss function, and the conservative N/A threshold (0.90) which prevents the model from incorrectly assigning a non-violence label when any evidence of violence is present.

Table 1
Label distribution in the Subtask 2 training set ($n = 4{,}845$).

Label	Violence Type	Positives	%
L0	Economic	890	18.4
L1	Physical	2,526	52.1
L2	Patrimonial	1,041	21.5
L3	Psychological	3,851	79.5
L4	Sexual	632	13.0
L5	Vicarious	533	11.0
L6	N/A	167	3.4

4. Methodology

Figure 1 gives an overview of the end-to-end system. Both subtasks share the same preprocessing, encoder, pooling, and multi-sample dropout stack; only the output head and loss function differ. We describe each stage in the order a report passes through it, from raw text to final prediction, and close

the section with a worked example (Section 4.7) that traces one real report from each subtask through the full pipeline.

4.1. Data Pre-Processing

The pipeline begins with the raw report text. Because this corpus consists of institutional case notes recorded by social workers (rather than web or social-media text), we first characterized it before designing the cleaning step. A scan of all 30,931 reports across both subtasks (train, development, and test) found no URLs and no user mentions, so the web-oriented normalization common in social-media pipelines is unnecessary here. Cleaning therefore reduces to whitespace normalization: runs of whitespace are collapsed to a single space, and leading and trailing whitespace is stripped. No lower-casing, lemmatization, or punctuation removal is applied, preserving the regional vocabulary and capitalization patterns present in the original reports. The cleaned text is then tokenized with the XLM-RoBERTa SentencePiece tokenizer and truncated to the subtask-specific maximum length (320 tokens for Subtask 1, 384 for Subtask 2); the effect of this truncation on the corpus is quantified in Section 9.

4.2. Model Architecture

We design a shared architecture (*BertMultiLabel*) built on XLM-RoBERTa-base [1], a 278M-parameter multilingual transformer pre-trained on 100 languages. It consists of three components.

Backbone. The XLM-RoBERTa-base encoder produces contextual token representations from the input sequence.

Attention-Masked Mean Pooling. Rather than relying on the `[CLS]` token, we aggregate information across all tokens using a weighted mean that excludes padding positions:

$$\mathbf{p} = \frac{\sum_t m_t \cdot \mathbf{h}_t}{\max(\sum_t m_t,\ \epsilon)} \tag{1}$$

where $\mathbf{h}_t$ is the hidden state at position t, $m_t \in \{0, 1\}$ is the attention mask, and $\epsilon = 10^{-6}$. This is especially suited to the long, narrative-style reports in this corpus, where relevant information is distributed across the full text.

Multi-Sample Dropout Classifier. Following Inoue [2], we instantiate $K{=}5$ independent dropout layers (each with $p{=}0.3$) sharing a single linear projection $\mathbf{W} \in \mathbb{R}^{768\times C}$, where C is the number of output classes. The final logits are the mean across all K samples:

$$\hat{\mathbf{y}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{W} \cdot \mathrm{Dropout}_k(\mathbf{p}) \tag{2}$$

This lightweight ensemble stabilizes gradients during training and reduces overfitting without requiring multiple model copies or increased inference cost beyond a single forward pass.

4.3. Common Training Configuration

Both subtasks share the following setup:

- **Optimizer:** AdamW [10] with Layer-wise Learning Rate Decay (LLRD). The top transformer layer uses a peak learning rate of 2×10^{-5}, each lower layer is scaled by a factor of 0.95 per layer reaching 5×10^{-6} at the embedding layer, and the classifier head is set $1.5\times$ higher than the top transformer layer (i.e., 3×10^{-5}).

- **LR Schedule:** Cosine annealing with 10% linear warmup.
- **R-Drop Regularization [4]:** Two independent forward passes are performed per mini-batch with different dropout masks. The symmetric KL divergence between the two output distributions is added to the primary loss with weight $\alpha = 0.1$.
- **Early stopping:** Based on development-set Macro F1-score.
- **Maximum epochs:** 20.

4.4. Subtask 1 Training Details

Table 2
Subtask 1 hyperparameters.

Hyperparameter	Value
Max sequence length	320 tokens
Micro batch size	12
Gradient accumulation	2 steps (effective batch 24)
Early stopping patience	4 epochs
Loss function	Focal Loss + label smoothing
Training labels	Hard labels

To address the severe class imbalance (Severe class $\approx 5\%$), we combine two complementary mechanisms. First, we use Focal Loss [3]:

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log p_t \tag{3}$$

with $\gamma = 2$ and per-class α_t weights inversely proportional to class frequency, applied with label smoothing $\varepsilon = 0.05$ to discourage overconfident predictions. Second, we use a `WeightedRandomSampler` during training with per-instance sampling weights inversely proportional to class frequency, so that each mini-batch contains a more balanced mixture of severity classes than the underlying training distribution. The system achieved a development Macro F1-score of 0.5056.

4.5. Subtask 2 Training Details

Table 3
Subtask 2 hyperparameters.

Hyperparameter	Value
Max sequence length	384 tokens
Micro batch size	12
Gradient accumulation	3 steps (effective batch 36)
Early stopping patience	5 epochs
Best checkpoint	Epoch 7 (stopped at Epoch 12)
Loss function	Weighted BCE + R-Drop
Training labels	Soft labels

Soft-label training. Instead of the binarized hard labels, we train on the raw mean annotation scores from the soft-label training file provided by the task organizers. These continuous values (e.g., 0.33, 0.67) encode annotator disagreement and teach the model to internalize the inherent ambiguity of GBV classification.

Loss function. The primary objective is Binary Cross-Entropy with Logits (`BCEWithLogitsLoss`) with per-label positive class weights:

$$w_j = \min\left(\frac{N - N_j^+}{N_j^+},\ 15\right) \tag{4}$$

where N is the total training size and N_j^+ is the count of positive instances for label j. The cap of $15\times$ prevents gradient instability on extremely rare labels. R-Drop regularization (KL divergence, $\alpha = 0.1$) is applied jointly. For this multi-label setting, the symmetric KL divergence is computed between the per-label sigmoid probabilities of the two forward passes and summed across labels.

4.6. Per-Label Threshold Optimization

After training, a grid search over candidate thresholds $\tau \in \{0.10, 0.15, \ldots, 0.90\}$ is run independently for each label on the development set:

$$\hat{\tau}_j = \arg\max_{\tau}\ F1_j(\tau) \tag{5}$$

The optimized thresholds are reported in Table 4.

Table 4
Per-label optimized decision thresholds for Subtask 2, selected by maximizing per-label F1 on the development split.

Label	Violence Type	Threshold $\hat{\tau}_j$
L0	Economic	0.65
L1	Physical	0.40
L2	Patrimonial	0.80
L3	Psychological	0.40
L4	Sexual	0.65
L5	Vicarious	0.80
L6	N/A	0.90

On the development split, per-label threshold optimization yielded an absolute improvement of $+0.017$ in Macro F1 over the default threshold of 0.50, without any additional training. The final per-label breakdown of precision, recall, and F1 with the optimized thresholds is reported in Section 6 (Table 9).

4.7. Worked Example: Tracing Two Reports Through the Pipeline

To make the end-to-end flow concrete, we trace one real report from each subtask through every stage of Figure 1, from raw text to final prediction.

Subtask 1 (severity). Consider a representative report from the test set whose raw text reads:

> *"USUARIA ASISTE A ASESORIA DE DIVORCIO Y PENSION, ESTAN SEPARADOS DESDE HACE TRES SEMANAS, Y DESDE ENTONCES NO LE DA MANUTENCION"*
>
> 'The user attends counseling for divorce and child support; the couple separated three weeks ago and has received no maintenance since.'

Because the text contains no URLs or user mentions, pre-processing only collapses and strips whitespace, leaving the original wording and capitalization intact. The cleaned text is tokenized into a short sequence of subword units (well within the 320-token limit) and padded to a fixed length, with an attention mask marking the real tokens. XLM-RoBERTa-base encodes this sequence into one 768-dimensional

contextual vector per token, and attention-masked mean pooling averages the non-padding vectors into a single 768-dimensional report representation. This vector passes through the multi-sample dropout head (a single deterministic pass at inference) and the shared linear layer, producing four severity logits; a softmax converts these into a distribution over {Mild, Medium, High, Severe}, and the *argmax* of that distribution is the predicted severity. For this report our submitted system predicts *Mild*, consistent with its purely administrative content: a divorce-and-support counseling request with no mention of physical, sexual, or escalating harm.

Subtask 2 (violence types). Now consider development report #12062, in which a former partner:

> *"LA SIGUE MOLESTANDO, LA QUIERE OBLIGAR A QUE REGRESE … CHANTAJEANDOLA CON EL MENOR QUE TIENEN EN COMUN … LA ACOSA CONSTANTEMENTE"*
>
> '… keeps harassing her and trying to force her to come back, blackmailing her with their shared child … and constantly stalks her.'

Pre-processing and tokenization proceed exactly as above, here against the 384-token limit, and the encoder and mean-pooling stages again yield a single 768-dimensional representation. Because Subtask 2 is multi-label, the shared linear layer emits seven independent logits (one per violence type) and a sigmoid maps each logit to an independent probability. Each probability is then compared against its own optimized threshold from Table 4: a label is emitted only when its probability exceeds that threshold. The report conveys sustained harassment, coercion, and emotional manipulation, but no physical, sexual, economic, or patrimonial component; accordingly, only the Psychological (L3) probability clears its threshold (0.40) while the remaining six fall below theirs, and the system outputs the single label *Psychological*, matching the gold annotation.

This example also illustrates why per-label thresholds matter: under a single global threshold of 0.50, the high-recall labels Physical and Psychological ($\hat{\tau} = 0.40$) would be harder to trigger, whereas rare labels such as N/A ($\hat{\tau} = 0.90$) rely on their stricter thresholds to suppress spurious activations.

5. Experimental Setup

All experiments use the Hugging Face Transformers library [11] with PyTorch [12], loading the tokenizer and pretrained weights for `xlm-roberta-base` directly from the Hugging Face model hub. The two subtasks were trained on separate machines: Subtask 1 on a workstation with an NVIDIA RTX 5050 GPU (8 GB VRAM), and Subtask 2 on a laptop with an NVIDIA RTX 3050 Ti GPU (4 GB VRAM). The full software stack for each environment, together with the configuration shared across both subtasks, is reported in Table 5.

The XLM-RoBERTa-base encoder and the task-specific classifier head are trained end-to-end from the pretrained checkpoint for each subtask independently. No weight sharing or joint training is performed between the two subtasks. The best checkpoint is selected based on the development set metric: Macro F1-score for Subtask 1, and Micro F1-score for Subtask 2. Final test predictions are generated exclusively from the selected best checkpoint without any ensembling or post-training modification beyond the threshold optimization described in Section 4.

6. Results

Official results on the withheld test set are presented in Tables 6 and 7.

Table 8 compares development and test performance across both subtasks. For Subtask 2, the development numbers are reported with the optimized per-label thresholds learned on the development split itself.

Our Subtask 1 test score (0.5565) exceeds our development score (0.5056). We attribute this gap mainly to differences between the development and test distributions rather than to a confirmed generalization gain, and note that our single-seed setup (Section 9) cannot establish such an effect with confidence.

Table 5
Software and hardware environments per subtask. The two subtasks were trained on different machines. The bottom block lists the configuration shared by both subtasks.

Component	Subtask 1	Subtask 2
GPU	NVIDIA RTX 5050 (8 GB)	NVIDIA RTX 3050 Ti (4 GB)
CPU	Intel Core i7-14650HX	AMD Ryzen 7 6800H
RAM	24 GB	16 GB
Python	3.14.2	3.11
PyTorch	2.11.0 (CUDA 12.8)	2.5.1 (CUDA 12.1)
Transformers	5.8.0	5.7.0
Shared across both subtasks:		
Base model	`xlm-roberta-base` (278M parameters)	
Optimizer	AdamW with Layer-wise LR Decay	
LR schedule	Cosine annealing + 10% warmup	
Early stopping	Development-set Macro F1	
Maximum epochs	20	

Table 6
Official Subtask 1 results (selected teams).

Team	Macro F1	Rank
angellabco	0.6272	1st
franrodrigo	0.6199	2nd
marcelhr	0.6080	3rd
manojkumar00011 (ours)	**0.5565**	**11th / 16**
ymlopez	0.4835	16th

Table 7
Official Subtask 2 results (selected teams).

Team	Micro F1	Hamming Loss	Rank
jorgegomezn	0.8879	0.0631	1st
verbanexai	0.8782	0.0692	2nd
phonghoang	0.8740	0.0711	3rd
manojkumar00011 (ours)	**0.8584**	**0.0818**	**12th / 15**
ymlopez	0.7689	0.1406	15th

For Subtask 2, dev and test Micro F1 are within 0.003 of each other (0.8609 vs. 0.8584), indicating that the per-label thresholds tuned on the development set transfer cleanly to the unseen test set. Our test Micro F1 of 0.8584 places us within 0.03 of the first-place system (0.8879), and well above the last-placed team (0.7689).

Table 9 reports per-label precision, recall, and F1 on the development set with the optimized thresholds. Frequent and linguistically explicit labels, Physical (L1) and Psychological (L3), achieve F1 above 0.91. Sexual (L4) violence also reaches a strong 0.867 despite its lower prevalence. The weakest label is Vicarious violence (L5, F1 = 0.555), which is conceptually more abstract (violence inflicted through harm to a third party such as a child) and is therefore harder to identify from surface lexical cues. Patrimonial (L2) and N/A (L6) sit in the middle range, with N/A showing the expected pattern of high recall (0.929) but lower precision (0.672) under its conservative threshold of 0.90.

Table 8
Development vs. test performance summary. Test Macro F1 for Subtask 2 is unavailable (—), as the official leaderboard reports only Micro F1 and Hamming Loss for that subtask.

Metric	Development	Test
S1 Macro F1	0.5056	**0.5565**
S2 Macro F1 (optimized τ)	0.8006	—
S2 Micro F1 (optimized τ)	0.8609	**0.8584**
S2 Hamming Loss	0.0805	**0.0818**

Table 9
Per-label precision, recall, and F1 on the Subtask 2 development set ($n = 1{,}212$) using the optimized thresholds from Table 4.

Label	Violence Type	Positives	Precision	Recall	F1
L0	Economic	244	0.759	0.914	0.829
L1	Physical	616	0.904	0.921	**0.912**
L2	Patrimonial	249	0.745	0.751	0.748
L3	Psychological	931	0.892	0.936	**0.913**
L4	Sexual	172	0.862	0.872	0.867
L5	Vicarious	119	0.490	0.639	0.555
L6	N/A	42	0.672	0.929	0.780
				Macro F1	0.801
				Micro F1	0.861

7. Discussion

Subtask 1. The primary bottleneck is the Severe class, which represents only $\approx 5\%$ of training samples. While Focal Loss partially mitigates this by up-weighting hard minority examples, the extreme imbalance limits the ceiling of Macro F1. The 0.07 gap between our score and first place (0.5565 vs. 0.6272) suggests that ensemble methods, pseudo-labeling on unlabeled data, or augmentation techniques targeting the Severe class could yield further gains.

Subtask 2. The combination of soft-label training, weighted BCE, and per-label threshold optimization proved effective. Soft labels communicate annotator uncertainty to the model. For instance, a sample scored 0.67 for Physical violence (two of three annotators agreed) is treated differently from one scored 1.0 (full agreement). The low optimized thresholds for Physical (L1=0.40) and Psychological (L3=0.40) violence reflect their high prevalence and frequent co-occurrence. The conservative N/A threshold (L6=0.90) prevents the model from incorrectly labeling violent cases as non-violent when it has even moderate evidence of violence.

The $+0.017$ gain from threshold optimization underscores the value of post-hoc calibration for multi-label tasks with imbalanced distributions: a simple grid search on the development set can meaningfully improve final performance without any additional training.

Architecture choices. Attention-masked mean pooling was chosen over `[CLS]`-based representations as it better captures information distributed across long narrative reports (many of which exceed 200 tokens) rather than relying solely on the first token. Multi-sample dropout and R-Drop were adopted to improve training stability, particularly for Subtask 2 where the training set is smaller (4,845 samples); we did not run controlled ablations isolating their individual contributions.

8. Conclusion

We presented a multilingual transformer system for the WomenHelp 2026 shared task that addresses the automatic classification of GBV reports in Mexican Spanish. Built on XLM-RoBERTa-base with multi-sample dropout and attention-masked mean pooling, our system uses task-specific training strategies: Focal Loss with label smoothing for multi-class severity classification (Subtask 1), and weighted BCE with soft-label training and per-label threshold optimization for multi-label violence type identification (Subtask 2). We achieved Macro F1 of 0.5565 (11^{th} of 16) on Subtask 1 and Micro F1 of 0.8584 with Hamming Loss of 0.0818 (12^{th} of 15) on Subtask 2.

Future work could explore: (1) ensembling multiple XLM-RoBERTa checkpoints or complementary multilingual models; (2) leveraging Spanish-domain pre-trained models such as RoBERTa-BNE; (3) using Subtask 1 severity labels as auxiliary supervision for Subtask 2; and (4) data augmentation targeting the underrepresented Severe class.

9. Limitations

Several limitations of the present work should be acknowledged. First, all experiments were carried out on consumer-grade GPUs (an 8 GB RTX 5050 for Subtask 1 and a 4 GB RTX 3050 Ti for Subtask 2). These memory budgets constrained both the maximum batch size and the maximum sequence length, and made full fine-tuning of larger XLM-RoBERTa variants (Large, XL) infeasible. The chosen length limits truncate a non-trivial minority of reports: 14.1% of Subtask 1 training reports (and 22.4% of its development set) exceed 320 tokens, while 10.0% of Subtask 2 training reports exceed 384 tokens, with the longest reports reaching roughly 750 subword tokens. Because violence cues can appear anywhere in these long narratives, this truncation may discard relevant evidence, particularly for Subtask 2, where several violence types may be distributed across the full report, and a larger context window is therefore a natural avenue for improvement. Second, our reported results correspond to a single random seed (42); a more rigorous evaluation would average across multiple seeds with confidence intervals, and we expect non-trivial seed variance, especially on the Severe class of Subtask 1 and on the rare labels of Subtask 2 (Vicarious, N/A). Third, per-label thresholds for Subtask 2 were tuned on the development split, which carries a small risk of overfitting to that split; the close agreement between dev Micro F1 (0.861) and test Micro F1 (0.858) suggests this risk did not materialize, but a held-out tuning split would be cleaner. Fourth, we did not perform a qualitative error analysis on the Severe class or on the Vicarious label, both of which show the weakest performance and would benefit from a closer linguistic study. Fifth, the system is trained and evaluated on a corpus from a specific regional and institutional context (Northern Mexico, women's support centers) and its generalization to other Spanish-speaking regions, registers, or to other forms of GBV documentation has not been established. Finally, we did not run controlled ablations isolating the contribution of each component (multi-sample dropout, R-Drop, soft-label training, and per-label positive weighting); the gains we attribute to them are therefore motivated by design rather than quantified, and isolating their individual effects is left to future work.

Code and Data Availability

The dataset is provided by the WomenHelp 2026 shared task organizers through the CodaBench platform.[1] The implementation, trained model checkpoints, prediction files, and threshold configurations supporting the results reported in this paper are available from the corresponding author on reasonable request.

[1] https://www.codabench.org/competitions/14614/

Acknowledgments

The authors thank the WomenHelp 2026 organizing committee for providing a challenging and socially meaningful shared task, and the Department of Computer Science and Engineering at Sri Sivasubramaniya Nadar College of Engineering for hosting the Summer Internship program under which this work was carried out. We are grateful to the expert annotators whose careful labeling of the corpus made this research possible.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. All experimental design, model training, data analysis, and reported results are the authors' own.

References

[1] 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.

[2] H. Inoue, Multi-sample dropout for accelerated training and better generalization, arXiv preprint arXiv:1905.09788 (2019).

[3] 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.

[4] Y. Liang, P. He, Y. Gong, N. Lin, L. Shou, D. Jiang, N. Duan, R-Drop: Regularized dropout for neural language models, in: Advances in Neural Information Processing Systems, volume 34, 2021, pp. 10890–10905.

[5] 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).

[6] 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.

[7] 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, Association for Computational Linguistics, 2019, pp. 4171–4186.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[9] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining 3 (2007) 1–13.

[10] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019.