IReL_IIT(BHU) at WomenHelp 2026: Transformer-Based Severity and Multi-Label Classification of Gender-Based Violence Reports

Arjun Mukherjee*,†, Sukomal Pal†

Indian Institute of Technology (Banaras Hindu University), Varanasi, India

Abstract

This paper describes our participation in the IberLEF 2026 WomenHelp-MX shared task on automatic classification of gender-based violence (GBV) reports from northern Mexico. The task presents two subtasks: (1) severity classification of reports into four levels (Mild, Medium, High, Severe), and (2) multi-label identification of seven violence types (Economic, Physical, Patrimonial, Psychological, Sexual, Vicarious, N/A). We fine-tune XLM-RoBERTa Large for Subtask 1 and BETO for Subtask 2 on the provided training data. Our system achieves a macro-F1 of **0.6058** in Subtask 1 (**4th place out of 16 teams**) and a micro-F1 of **0.8670** in Subtask 2 (**9th place out of 15 teams**). Analysis highlights systematic under-prediction of the Severe class across all submitted systems, and low recall for the Vicarious violence type, both linked to class imbalance and genuine annotator disagreement in the training data.

Keywords

gender-based violence, text classification, multi-label classification, transformers, BETO, XLM-RoBERTa Large, IberLEF 2026

1. Introduction

Violence against women is a pervasive global problem with physical, psychological, economic, sexual, patrimonial, and vicarious manifestations. Automating the analysis of reports documenting gender-based violence (GBV) offers significant potential to support social workers, legal professionals, and policymakers by enabling systematic processing of large case volumes that would otherwise require extensive manual effort.

The IberLEF 2026 WomenHelp-MX shared task [1], part of the IberLEF 2026 evaluation forum [2], presents a unique and socially valuable challenge: classifying real, anonymised GBV reports from northern Mexico. The task demands systems capable of processing the formal-institutional register of social workers alongside the regional lexical variation of northern Mexican Spanish—a combination not well represented in standard NLP benchmarks.

The task comprises two subtasks. **Subtask 1** is a four-class severity classification problem (Mild, Medium, High, Severe). **Subtask 2** is a seven-label multi-label classification problem (Economic, Physical, Patrimonial, Psychological, Sexual, Vicarious, N/A). Both subtasks are evaluated primarily by macro-F1, which places equal weight on all classes—a deliberate and appropriate choice that emphasises performance on rare, high-stakes categories.

We describe our approach, which fine-tunes pre-trained transformer models (XLM-RoBERTa Large for Subtask 1, BETO for Subtask 2) on the provided training data. Our system ranked **4th of 16 teams** in Subtask 1 (macro-F1: 0.6058) and **9th of 15 teams** in Subtask 2 (micro-F1: 0.8670).

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

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

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

2. Related Work

We survey prior work across three areas directly relevant to our approach: NLP for gender-based violence detection, transformer models for Spanish text classification, and multi-label learning.

2.1. NLP for Gender-Based Violence Detection

The application of NLP to gender-based violence has attracted growing attention. Early work focused on hate speech and misogyny detection on social media [3, 4]. A comprehensive systematic review by Abercrombie et al. [5] surveyed 63 resources for automated identification of online GBV, finding that datasets are limited by lack of theoretical grounding and stakeholder input—observations that underscore the value of expert-annotated corpora such as WomenHelp-MX. Shared tasks have been instrumental in advancing the field: the EXIST series on sexism identification in social networks [6, 7] has run continuously since 2021, evolving to address multimodal content and learning-with-disagreement paradigms. WomenHelp-MX [1] is distinguished by its focus on real victim reports rather than social media, its explicit severity grading, and its multi-label violence type structure—none of which have been addressed together in prior shared tasks.

2.2. Transformer Models for Spanish Text Classification

BETO [8] is a BERT-based model pre-trained on a large Spanish corpus using whole-word masking. XLM-RoBERTa [9] is a multilingual model trained on 100 languages, showing competitive or superior performance to monolingual models on cross-domain tasks.

2.3. Multi-label Text Classification

Multi-label classification—where each instance may belong to multiple categories simultaneously—is well studied [10]. Transformer-based approaches typically frame the problem as independent binary classification per label with sigmoid activations and binary cross-entropy loss. Soft label training [11] has been proposed to handle annotator disagreement, which is present in the WomenHelp-MX dataset through the provided inter-annotator soft labels.

3. Dataset

The WomenHelp-MX corpus consists of anonymised narratives from real GBV reports produced in northern Mexico. All identifying information (names, locations, dates) was removed by hand. The corpus preserves the institutional language of social workers and the regional lexical variation of northern Mexico.

3.1. Data Splits

Table 1 provides an overview of the dataset partitions. For Subtask 2, soft labels are also provided, containing mean annotation values across three expert social worker annotators (possible values: 0.0, 0.333, 0.667, 1.0). Hard binary labels use a threshold of 0.667 (two-thirds majority agreement).

3.2. Subtask 1: Severity Classes

The dataset exhibits notable class imbalance: the Severe class constitutes only approximately 4.5% of instances, while High represents approximately 39%. Exploratory analysis revealed a positive correlation between report length and severity (median word count: Mild = 50 words, Severe = 143 words), reflecting the more complex events described in serious cases. Table 2 and Figure 1 summarise the four classes and their distributional properties.

Table 1
Dataset partition overview.

Partition	Size	Labels
Task 1 — Train	13,630	Severity class (0–3)
Task 1 — Dev	3,405	Severity class (0–3)
Task 1 — Test	3,620	Blind (no labels)
Task 2 — Train	4,845	Multi-label hard + soft
Task 2 — Dev	1,212	Multi-label hard + soft
Task 2 — Test	3,620	Blind (no labels)

Table 2
Severity class distribution (Subtask 1).

Class	Description	Proportion
Mild (0)	Brief administrative / single minor incident	≈22%
Medium (1)	Moderate conflict; single incident reports	≈35%
High (2)	Repeated or multi-type violence	≈39%
Severe (3)	Life-threatening; extreme abuse	≈4.5%

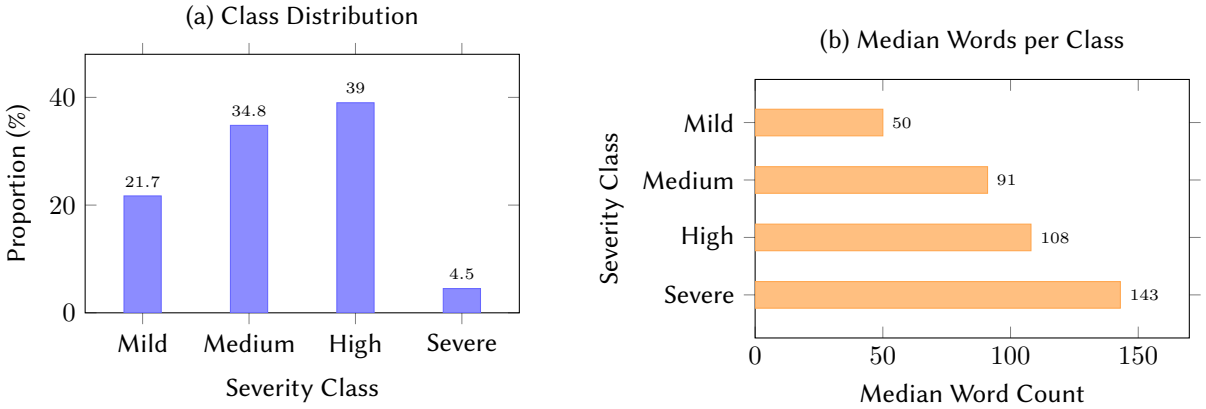

Figure 1: Subtask 1 dataset characteristics: (a) class distribution showing the severe imbalance of the Severe class (≈4.5%); (b) median word count per class, illustrating that report length correlates positively with severity.

3.3. Subtask 2: Violence Type Labels

Each report in Subtask 2 may carry one or more of seven labels (Table 3 and Figure 2). Psychological violence is near-ubiquitous (≈80%). The N/A label—indicating absence of explicit violence—constitutes only 3.4% of cases and is mutually exclusive with all other labels. The mean number of labels per document is approximately 2.0.

Vicarious violence shows the highest inter-annotator disagreement: it is the only label where minority-opinion annotations (soft value 0.333, one annotator agreeing) outnumber majority-opinion annotations (0.667, two annotators agreeing), indicating genuine conceptual ambiguity even among trained expert annotators.

4. Experimental Setup and Methodology

All experiments share a common pipeline of preprocessing, transformer encoding, and task-specific fine-tuning where the two subtasks differ primarily in model choice and training objective, as detailed in the subsections below.

Table 3
Subtask 2 label distribution (training set).

ID	Label	Prevalence	Train Count
L0	Economic	≈18%	890
L1	Physical	≈52%	2,526
L2	Patrimonial	≈22%	1,041
L3	Psychological	≈80%	3,851
L4	Sexual	≈13%	632
L5	Vicarious	≈11%	533
L6	N/A	≈3.4%	167

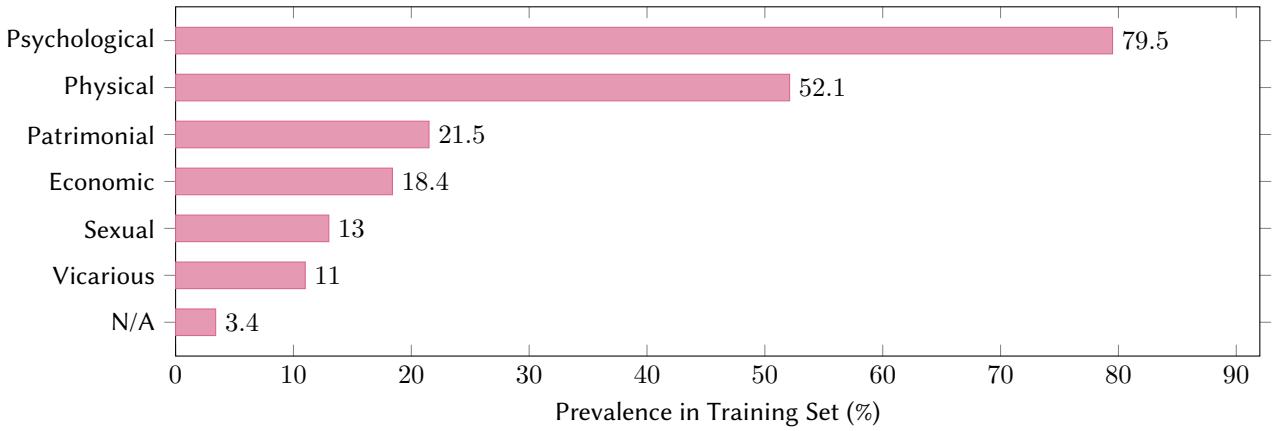

Figure 2: Subtask 2 label prevalence in the training set. Psychological violence dominates (≈80%) while Vicarious and N/A are rare, creating a severe multi-label imbalance that directly penalises macro-F1.

4.1. Preprocessing

Two preprocessing pipelines were applied depending on the system.

Baseline pipeline (FastText + XGBoost): Raw report text underwent the following sequence: (1) lowercasing; (2) punctuation removal; (3) numeral removal; and (4) lemmatisation using spaCy[1] with its Spanish model. The resulting lemmatised tokens were used to compute document-level embeddings.

Transformer pipeline (BETO, XLM-RoBERTa Large): Reports were normalised to standard Unicode and extraneous whitespace removed. Original casing was preserved, as both BETO and XLM-RoBERTa Large use cased tokenisation. No lemmatisation was applied.

4.2. System Overview

Figure 3 presents the end-to-end methodology applied across both subtasks.

4.3. Subtask 1: Severity Classification

We evaluated three systems for Subtask 1:

- **Baseline (FastText + XGBoost):** Text was preprocessed using the baseline pipeline described above. Document embeddings were produced by mean-pooling token-level FastText vectors from the Spanish Unannotated Corpora (SUC) [12]. A multi-class XGBoost classifier was trained on these embeddings with the following hyperparameters: n_estimators=500, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, gamma=1, reg_alpha=0.5, reg_lambda=1, min_child_weight=5.

[1] https://spacy.io

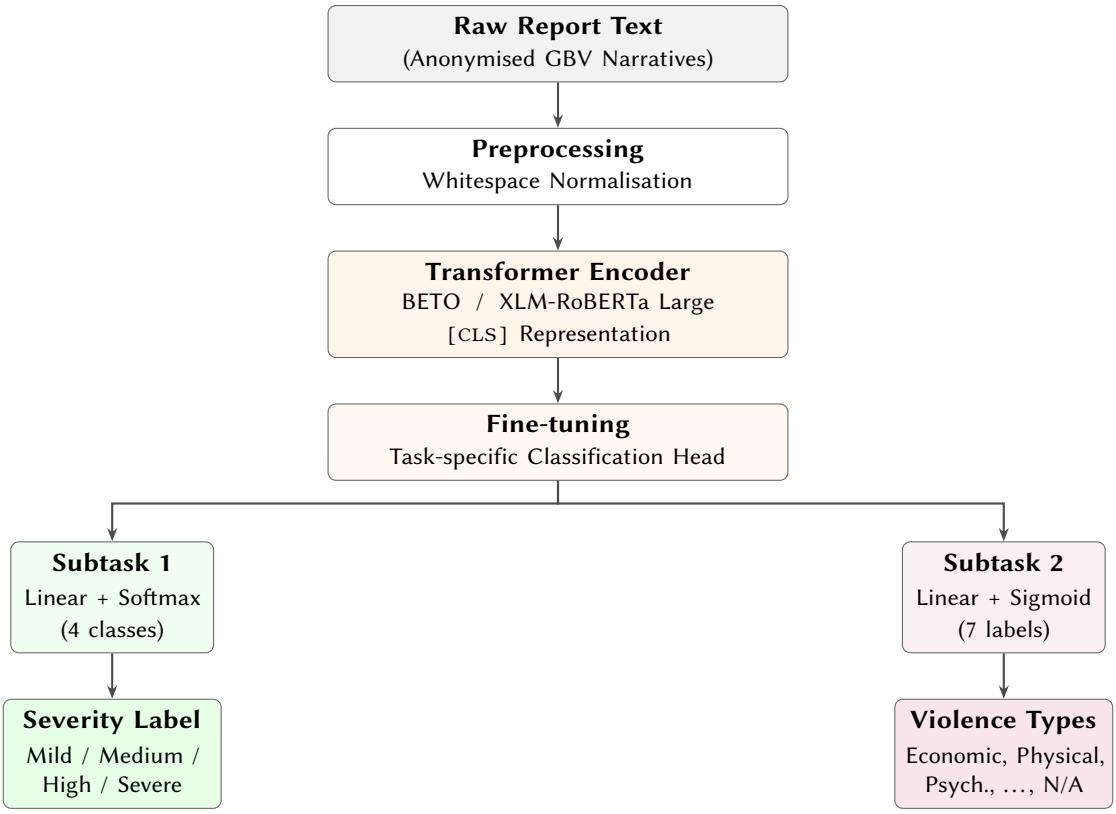

Figure 3: End-to-end system overview. A transformer encoder (BETO or XLM-RoBERTa Large) produces a `[CLS]` representation that is passed to a task-specific head: a Softmax head for Subtask 1 severity classification and a Sigmoid head for Subtask 2 multi-label violence type classification.

- **BETO:** The `dccuchile/bert-base-spanish-wwm-cased` model fine-tuned for four-class sequence classification. Preprocessing consisted of whitespace normalisation only (no lowercasing or lemmatisation). Training ran for up to 8 epochs with the best checkpoint selected by validation macro-F1.
- **XLM-RoBERTa Large (submitted):** The `FacebookAI/xlm-roberta-large` model fine-tuned for four-class sequence classification. Preprocessing was whitespace normalisation only, consistent with the BETO pipeline. Early stopping on validation loss was applied; training halted at Epoch 6 with the best checkpoint selected at Epoch 4 (macro-F1: 0.5076 on dev).

All transformer models used maximum sequence length 512, batch size 16, AdamW optimiser with learning rate 2e-5 and linear warmup, and up to 8 training epochs.

4.4. Subtask 2: Violence Type Classification

We fine-tuned BETO (`dccuchile/bert-base-spanish-wwm-cased`) with a multi-label classification head: a linear projection over the `[CLS]` representation followed by seven independent sigmoid outputs. `BCEWithLogitsLoss` (binary cross-entropy with logits) was used as the loss function. Training configuration: AdamW optimiser with linear learning rate schedule; up to 8 epochs; early stopping triggered after 2 consecutive non-improving validation losses, with the best checkpoint selected at Epoch 3 (lowest validation loss: 0.2112); batch size 16; decision threshold of 0.5 applied independently per label at inference. All four evaluation metrics (macro-F1, micro-F1, weighted F1, Hamming loss) were monitored per epoch. Soft labels were available as auxiliary data but were not incorporated into the final submission.

5. Evaluation Metrics

Systems are ranked by **macro-F1** for Subtask 1 and by **micro-F1** for Subtask 2. Additional metrics for Subtask 2 include weighted F1 and Hamming loss.

Macro-F1 is computed by averaging per-class F1 scores with equal weight across all classes. Under class imbalance this penalises equally for poor performance on rare classes (Severe, Vicarious) and common ones (High, Psychological), making it particularly demanding.

Micro-F1 aggregates true positives, false positives, and false negatives across all labels before computing precision and recall, and is therefore dominated by frequent labels.

Hamming Loss measures the fraction of label positions in the multi-hot prediction vector that are incorrect.

6. Results

We report results for both subtasks in turn, beginning with a system-by-system progression for Subtask 1 and followed by per-label and generalisation analysis for Subtask 2.

6.1. Subtask 1: Severity Classification

6.1.1. Baseline System

Table 4 presents the full per-class classification report for the FastText + XGBoost baseline evaluated on the development set (3,405 reports). The baseline achieves an overall accuracy of 53% and a macro-F1 of **0.4238**.

Table 4
Baseline (FastText + XGBoost) classification report on the development set.

Class	Precision	Recall	F1	Support
Mild (0)	0.57	0.48	0.52	739
Medium (1)	0.45	0.46	0.45	1,183
High (2)	0.58	0.66	0.62	1,329
Severe (3)	0.28	0.06	0.11	154
Accuracy			0.53	3,405
Macro avg	0.47	0.42	0.42	3,405
Weighted avg	0.52	0.53	0.52	3,405

The baseline reveals two fundamental failure modes that motivate our transformer-based approach. First, the Severe class achieves an F1 of only 0.11 — a recall of 0.06 means the model correctly identifies just 6% of Severe reports. This reflects the combination of extreme class imbalance (4.5% of the dataset) and the inability of mean-pooled static embeddings to capture the contextual cues distinguishing Severe from High violence. Second, Medium (F1 = 0.45) is poorly separated from the adjacent classes, consistent with the medium absorption bias observed across all systems.

6.1.2. BETO System

Table 5 shows the epoch-by-epoch training dynamics for the BETO system on the development set. Validation loss diverges from training loss from Epoch 1 onward, confirming progressive overfitting. The best macro-F1 is achieved at **Epoch 3** (0.4806), after which performance plateaus and early stopping halts training at Epoch 5.

The full per-class report at Epoch 3 (Table 6) shows substantial improvement over the baseline, particularly for Severe: recall rises from 0.06 to 0.26 and F1 from 0.11 to 0.31, confirming that contextual representations capture severity cues that static embeddings cannot. High remains the strongest class (F1 = 0.66) and Medium the hardest to separate (F1 = 0.48).

Table 5
BETO training dynamics on the development set (Subtask 1). Best checkpoint in bold.

Epoch	Train Loss	Val Loss	Macro-F1
1	0.7156	1.1313	0.4702
2	0.6103	1.2250	0.4346
3	**0.4819**	**1.3697**	**0.4806**
4	0.4322	1.5735	0.4740
5	0.3761	1.8420	0.4711

Table 6
BETO classification report on the development set (Epoch 3, best checkpoint).

Class	Precision	Recall	F1	Support
Mild (0)	0.59	0.39	0.47	739
Medium (1)	0.46	0.51	0.48	1,183
High (2)	0.62	0.70	0.66	1,329
Severe (3)	0.38	0.26	0.31	154
Accuracy			0.55	3,405
Macro avg	0.51	0.47	0.48	3,405
Weighted avg	0.55	0.55	0.54	3,405

6.1.3. XLM-RoBERTa Large System

Table 7 shows the training dynamics for XLM-RoBERTa Large. Validation loss again diverges from Epoch 0 onward while training loss decreases steadily. The best macro-F1 of **0.5076** is achieved at Epoch 4; a sharp validation loss spike at Epoch 5 (2.09) and Epoch 6 (2.56) triggers early stopping.

Table 7
XLM-RoBERTa Large training dynamics on the development set (Subtask 1). Best checkpoint in bold.

Epoch	Train Loss	Val Loss	Macro-F1
0	0.8168	1.1930	0.4530
1	0.6761	1.1033	0.3999
2	0.5748	1.3308	0.5001
4	**0.4365**	**1.5466**	**0.5076**
5	0.4341	2.0853	0.4951
6	0.3542	2.5565	0.4955

The per-class report at the best checkpoint (Epoch 4) is shown in Table 8. XLM-RoBERTa Large achieves the strongest Severe performance across all dev-set experiments: recall 0.31 and F1 = 0.36, compared to 0.31 for BETO and 0.11 for the baseline. Mild precision (0.64) is also the highest observed, reflecting improved class boundary learning from the larger multilingual model.

6.1.4. System Comparison

Table 9 summarises all three systems on the development set alongside the final test-set macro-F1 for our submitted system. XLM-RoBERTa Large achieves the best dev-set macro-F1 among all experiments (0.5076) and the best final test-set macro-F1 (**0.6058**), a gain of **+0.1820** over the FastText baseline and **+0.1252** over BETO. Each successive system improves Severe F1: 0.11 (baseline) $\rightarrow$ 0.31 (BETO) $\rightarrow$ 0.36 (XLM-RoBERTa Large), demonstrating that larger, contextual models are essential for the minority class that drives macro-F1.

Table 8
XLM-RoBERTa Large classification report on the development set (best checkpoint, Epoch 4).

Class	Precision	Recall	F1	Support
Mild (0)	0.64	0.51	0.56	739
Medium (1)	0.47	0.44	0.45	1,183
High (2)	0.60	0.72	0.66	1,329
Severe (3)	0.43	0.31	0.36	154
Accuracy			0.56	3,405
Macro avg	0.53	0.49	0.51	3,405
Weighted avg	0.55	0.56	0.55	3,405

Table 9
Subtask 1 system comparison. Dev-set scores marked †; bold = submitted system with final test-set score.

System	$F1_{Mild}$	$F1_{Med}$	$F1_{High}$	$F1_{Sev}$	Macro-F1
Baseline (FT+XGB)†	0.52	0.45	0.62	0.11	0.4238
BETO†	0.47	0.48	0.66	0.31	0.4806
XLM-RoBERTa Large†	0.56	0.45	0.66	0.36	0.5076
XLM-RoBERTa Large (test)	—	—	—	—	**0.6058**

†Evaluated on development set (3,405 reports).

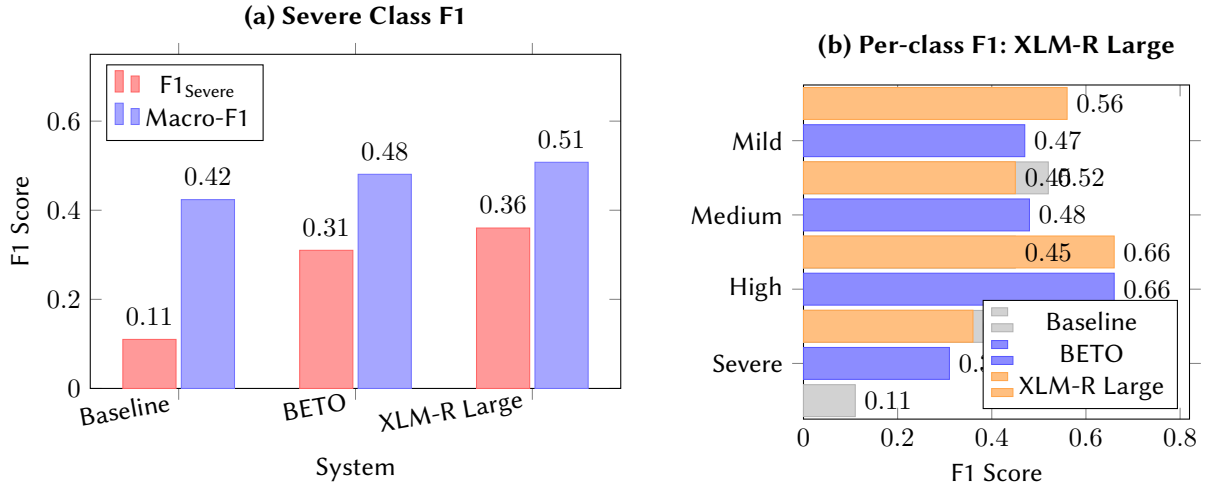

Figure 4: Our exclusive contribution to Subtask 1: progressive Severe class recovery across our three systems. (a) Each system improves both Severe F1 (0.11 → 0.31 → 0.36) and overall macro-F1. (b) Full per-class F1 comparison showing that XLM-RoBERTa Large improves Severe while maintaining competitive High performance.

6.2. Subtask 2: Violence Type Classification

6.2.1. Training Dynamics

Table 10 shows all four metrics across the five training epochs. Validation loss reaches its minimum at Epoch 3 (0.2112), triggering best-checkpoint selection. Unlike Subtask 1, macro-F1 continues to improve slightly beyond the best-loss checkpoint (0.7905 at Epoch 3 vs. 0.7951 at Epoch 5), confirming that validation-loss-based early stopping is suboptimal when macro-F1 is the target metric.

6.2.2. Per-label Performance

Table 11 presents the full per-label classification report at the best checkpoint (Epoch 3). Physical (F1 = 0.92) and Psychological (F1 = 0.90) are the strongest labels, benefiting from high support and clear

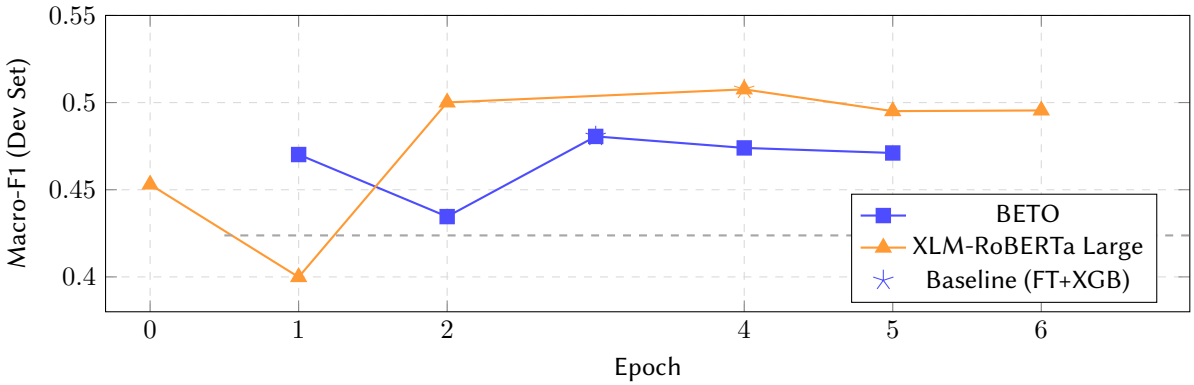

Figure 5: Subtask 1 macro-F1 on the development set across training epochs for BETO and XLM-RoBERTa Large. Stars (★) mark the best checkpoint per model. The dashed line shows the FastText+XGBoost baseline. BETO runs Epochs 1–5; XLM-RoBERTa Large runs Epochs 0,1,2,4,5,6 (Epoch 3 was not logged in the output).

Table 10
BETO training dynamics on the development set (Subtask 2). Best checkpoint (lowest val loss) in bold.

Ep.	Train L.	Val L.	Mac-F1	Mic-F1	Wt-F1	Hamming
1	0.3032	0.2920	0.5849	0.7959	0.7518	0.1087
2	0.2156	0.2408	0.7580	0.8349	0.8321	0.0950
3	**0.1778**	**0.2112**	**0.7905**	**0.8556**	**0.8529**	**0.0813**
4	0.1362	0.2163	0.7936	0.8566	0.8565	0.0814
5	0.1120	0.2294	0.7951	0.8555	0.8544	0.0810

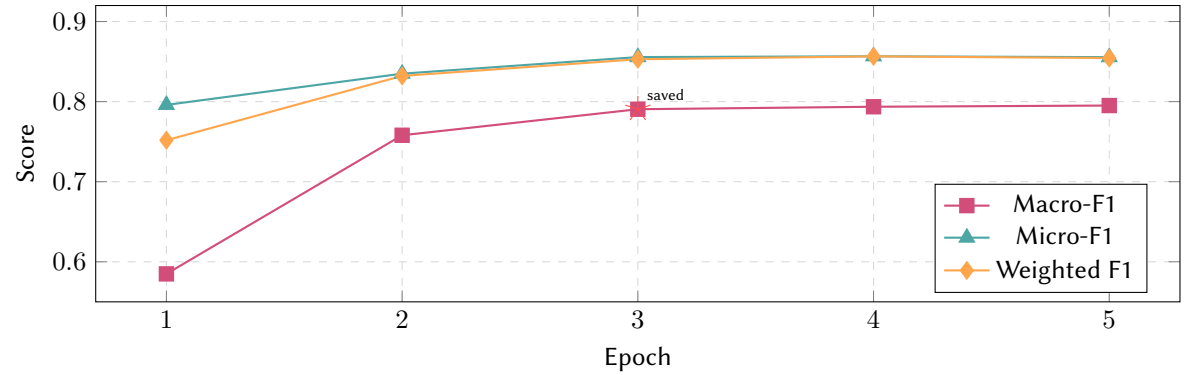

Figure 6: Subtask 2 BETO training curves across 5 epochs. The model is saved at Epoch 3 (lowest validation loss, marked ★) though macro-F1 continues improving through Epoch 5, illustrating the misalignment between validation loss and the primary evaluation metric.

lexical markers. Vicarious (F1 = 0.61) is the weakest label, driven by low recall (0.54), consistent with our earlier finding that Vicarious has the highest inter-annotator disagreement among all labels. Patrimonial (F1 = 0.73) also underperforms due to semantic overlap with both Economic and Physical violence.

6.2.3. Dev vs. Test Generalisation

Table 12 compares development and final test-set metrics. Micro-F1 improves from 0.8556 (dev) to **0.8670** (test) and Hamming loss improves from 0.0813 to 0.0753, indicating strong generalisation without overfitting to the development set.

Table 11
BETO per-label classification report on the development set (Epoch 3, best checkpoint). Saved macro-F1: 0.7905.

Label	Precision	Recall	F1	Support
Economic	0.82	0.79	0.81	244
Physical	0.90	0.93	0.92	616
Patrimonial	0.75	0.71	0.73	249
Psychological	0.87	0.93	0.90	931
Sexual	0.80	0.80	0.80	172
Vicarious	0.70	0.54	0.61	119
N/A	0.82	0.74	0.78	42
Micro avg	0.85	0.86	0.86	2373
Macro avg	0.81	0.78	0.79	2373
Weighted avg	0.85	0.86	0.85	2373
Samples avg	0.86	0.88	0.85	2373

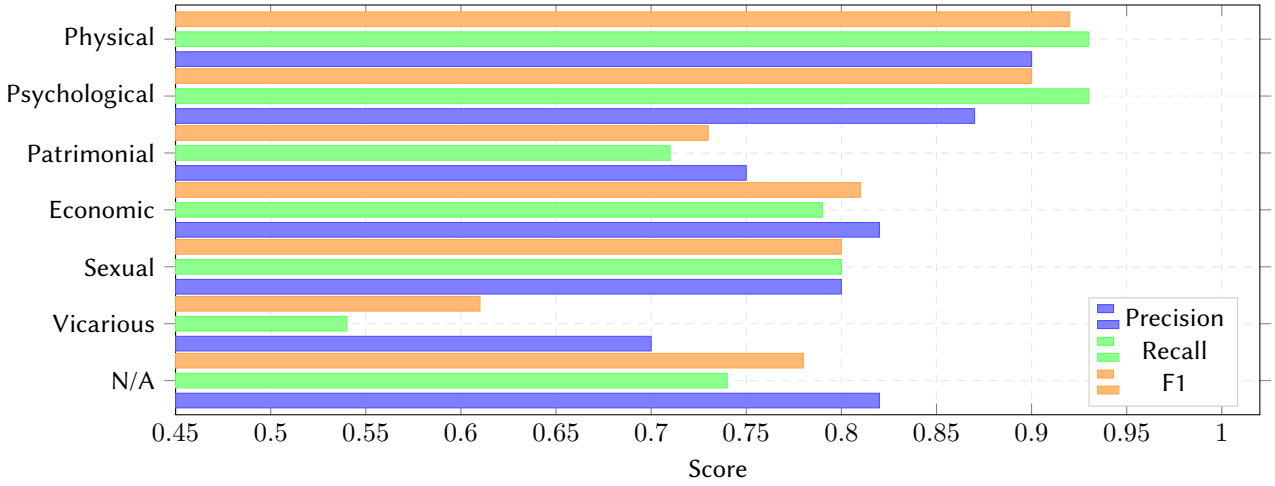

Figure 7: Subtask 2 per-label precision, recall, and F1 for BETO at the best checkpoint (Epoch 3). Vicarious has the largest precision–recall gap (0.70 vs. 0.54), confirming that the model finds Vicarious when confident but misses many true cases.

Table 12
Subtask 2 overall metrics: development set (Epoch 3) vs. final test set.

Metric	Dev (Ep. 3)	Test	Change
Macro-F1	0.7905	— (not ranked)	—
Micro-F1	0.8556	**0.8670**	+0.0114
Weighted F1	0.8529	—	—
Hamming Loss	0.0813	**0.0753**	−0.0060

7. Competition Standings

Tables 13 and 14 present the final leaderboards for both subtasks. Our system (IReL_IIT(BHU)) is marked with (⋆).

Our system ranked 4th of 16 teams in Subtask 1 with a macro-F1 of 0.6058, trailing the top-performing team (angellabco, 0.6272) by **0.0214**. In Subtask 2 we ranked 9th of 15 teams with a micro-F1 of 0.8670, trailing the top-performing team (jorgegomezn, 0.8879) by **0.0209**. A cross-task analysis reveals that the two subtasks favour different modelling strengths: the Subtask 1 top team placed only 8th in Subtask 2, while the Subtask 2 top team placed only 6th in Subtask 1, suggesting the tasks reward distinct capabilities.

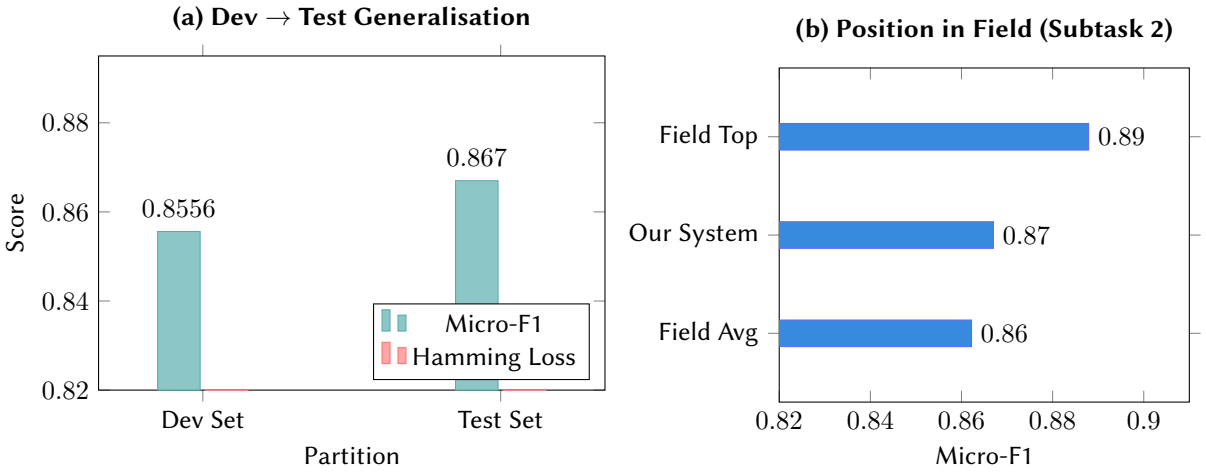

Figure 8: Our exclusive contribution to Subtask 2: (a) Dev-to-test generalisation — micro-F1 *improves* from dev (0.8556) to unseen test (0.8670) and Hamming loss decreases (0.0813 → 0.0753), confirming no overfitting despite the test set being 60% larger than combined train+dev; (b) our position relative to the field mean (0.8622) and top-performing system (0.8879).

Table 13
Final leaderboard — Subtask 1: Severity Classification (Macro-F1). ⋆ = our system.

Place	Team	Macro-F1	Gap to Top
1st	angellabco	0.6272	—
2nd	franrodrigo	0.6199	−0.0073
3rd	marcelhr	0.6080	−0.0192
4th ⋆	**IReL_IIT(BHU)**	**0.6058**	**−0.0214**
5th	omargarcia	0.5936	−0.0336
6th	jorgegomezn	0.5915	−0.0356
7th	aerojash	0.5750	−0.0522
8th	aalgarra	0.5723	−0.0548
9th	phonghoang	0.5713	−0.0558
10th	danielrod	0.5675	−0.0596
11th–16th	…	0.4835–0.5565	—

8. Error Analysis

We analyse systematic error patterns across both subtasks, focusing on class imbalance effects in Subtask 1 and label-level recall failures and training objective misalignment in Subtask 2.

8.1. Severe Class Under-prediction (Subtask 1)

The most consequential error pattern across all submitted systems is the systematic under-prediction of the Severe class. The baseline makes this starkest: Severe recall is only 0.06, meaning 94% of actual Severe reports are missed, yielding F1 = 0.11 despite precision 0.28. BETO substantially improves Severe recall to 0.26 (F1 rising to 0.31), but still misses nearly three quarters of Severe cases. Our XLM-RoBERTa Large system further improves on this, yet all systems systematically under-predict Severe relative to its true prevalence. A consistent *medium absorption bias* is observed across all systems: uncertain cases are assigned to Medium rather than the extremes, as expected when cross-entropy loss is applied with uniform class weights over an imbalanced distribution.

Text length provides a strong signal for the extremes (median word count: Mild = 50, Severe = 143), but the semantic distinction between High and Severe rests on fine-grained details—repeated incidents, presence of weapons, explicit life threats—that mean-pooled static embeddings cannot capture and

Table 14
Final leaderboard — Subtask 2: Violence Type Classification (Micro-F1 and Hamming Loss). ⋆ = our system.

Place	Team	Micro-F1	Hamming	Gap to Top
1st	jorgegomezn	0.8879	0.0631	—
2nd	verbanexai	0.8782	0.0692	−0.0097
3rd	phonghoang	0.8740	0.0711	−0.0139
4th	aerojash	0.8735	0.0721	−0.0143
5th	pln_uam	0.8715	0.0749	−0.0164
6th	marcelhr	0.8702	0.0758	−0.0177
7th	franrodrigo	0.8700	0.0731	−0.0178
8th	angellabco	0.8677	0.0767	−0.0202
9th ⋆	**IReL_IIT(BHU)**	**0.8670**	**0.0753**	**−0.0209**
10th	aalgarra	0.8669	0.0742	−0.0209
11th–15th	…	0.7689–0.8665	0.0766–0.1406	—

that are challenging even for contextual models. The BETO training curve (Table 5) further shows that validation loss diverges from training loss from Epoch 1, yet macro-F1 peaks at Epoch 3 rather than at stopping, confirming that early stopping on validation loss halts training before the model has fully learned minority class boundaries.

8.2. Vicarious and Patrimonial Labels (Subtask 2)

Vicarious is the label with the highest error rate (F1 = 0.61), driven by low recall (0.54). Soft-label analysis confirms that Vicarious has the highest inter-annotator disagreement in training data—the only label where minority-opinion annotations (0.333) outnumber majority-opinion ones (0.667)—indicating genuine conceptual difficulty.

Patrimonial (F1 = 0.73) is precision-heavy (precision 0.75, recall 0.71), reflecting the model's conservatism on a label that overlaps semantically with both Economic and Physical violence. The uniform decision threshold of 0.5 is suboptimal; per-label threshold tuning on the development set—lowering the Vicarious threshold to approximately 0.30–0.35—would be expected to yield material macro-F1 gains.

8.3. Early Stopping Misalignment (Subtask 2)

The model was saved at Epoch 3 (validation loss: 0.2112, the minimum) despite macro-F1 continuing to improve through Epoch 5 (0.7905 at Epoch 3 vs. 0.7951 at Epoch 5, a difference of +0.0046). The training curve (Table 10) makes this divergence explicit: validation loss rises monotonically after Epoch 3 while macro-F1 keeps improving, with the best Hamming loss (0.0810) also at Epoch 5. Early stopping with patience of 2 on validation loss correctly prevents severe overfitting, but selects a checkpoint that is suboptimal for the ranking metric. Future work should monitor macro-F1 directly as the early stopping criterion, or use a composite signal combining loss stabilisation and F1 improvement.

9. Conclusion

We presented our system for the IberLEF 2026 WomenHelp-MX shared task. Fine-tuning pre-trained transformer models (XLM-RoBERTa Large for Subtask 1, BETO for Subtask 2) yielded competitive results: 4th of 16 teams in Subtask 1 (macro-F1: 0.6058) and 9th of 15 in Subtask 2 (micro-F1: 0.8670). Three primary limitations were identified: (1) the Severe severity class is systematically under-predicted by all competing systems; focal loss or class-weighted training would be the natural remedy. (2) Vicarious violence suffers from genuinely ambiguous annotation boundaries; soft label training and per-label threshold tuning would address this. (3) Using validation loss as an early stopping criterion in a macro-F1-optimised task leads to premature model selection and should be replaced by direct macro-F1 monitoring. The WomenHelp-MX dataset represents an important and underserved domain for NLP,

and this work contributes to the broader goal of developing tools that support social workers and justice professionals.

Acknowledgments

The authors gratefully acknowledge Krishna Tewari for her valuable insights and constructive guidance at every stage of this work, from participation through to final submission. The authors also thank the WomenHelp 2026 shared task organisers for curating and making available this valuable corpus, and for their continued support throughout the competition.

Declaration on Generative AI

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

References

[1] 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] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at IberEval, in: Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR-WS.org, 2018.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.

[5] G. Abercrombie, A. Jiang, P. Gerrard-Abbott, I. Konstas, V. Rieser, Resources for automated identification of online gender-based violence: A systematic review, in: Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 170–186.

[6] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism identification in social neTworks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[7] L. Plaza, J. Carrillo-de Albornoz, F. Rodríguez-Sánchez, J. Gonzalo, P. Rosso, D. Spina, Overview of EXIST 2024 — learning with disagreement for sexism identification and characterization in tweets and memes, in: Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer, 2024.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Fasoli, M. Auli, 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.

[10] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Transactions on Knowledge and Data Engineering 26 (2014) 1819–1837.
[11] R. Müller, S. Kornblith, G. E. Hinton, When does label smoothing help?, in: Advances in Neural Information Processing Systems, volume 32, 2019.
[12] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.