WH_comp_team at WomenHelp 2026: MoECORN for Degrees of Violence Classification and ModernBERT Based Label Wise Attention for Type of Violence Detection

Alberto Algarra[1,*,†], Sofía Jimeno[2,†]

[1]BBVA AI Factory, BBVA, Madrid, Spain
[2]Independent researcher, Porto, Portugal

Abstract

This paper presents two computationally efficient approaches that require limited training resources to tackle the tasks proposed in the WomenHelp 2026 shared task: *Classification of Gender-Based Violence Reports to Support Threatened Women in Northern Mexico.* Specifically, we address (i) the classification of the severity level of violence in judicial reports and (ii) the identification of the types of violence described in these reports, while prioritizing efficiency and simplicity over large-scale architectures.

For the first subtask, we propose a Mixture of Experts architecture that leverages sentence transformer representations from models trained on diverse domains, combined with a CORN (Conditional Ordinal Regression for Neural Networks) loss to capture the ordinal nature of violence severity. Despite its simplicity and low computational cost, our approach achieves a micro-F1 score of 0.59 and a macro-F1 score of 0.57 in the blind test set, remaining competitive with more complex alternatives.

For the second subtask, we introduce an efficient soft-label multi-label classification model based on a ModernBERT encoder pretrained on Spanish and English corpora. The model incorporates a label-wise attention mechanism with independent outputs per label and is trained using a composite loss function that combines binary cross-entropy with Jensen–Shannon divergence. Without relying on large-scale or heavily engineered architectures, this approach achieves a macro-F1 of 0.79, micro-F1 of 0.87, weighted F1 of 0.87, and a Hamming loss of 0.07, demonstrating strong performance relative to more resource-intensive methods.

Keywords

Violence against women, Mixture of experts, Ordinal classification, Label-wise attention, Sentence transformers, ModernBERT

1. Introduction

Violence against women is a pervasive social problem with a significant impact across diverse cultural and geographical contexts. Its manifestations extend beyond physical aggression to include psychological, sexual, economic, and other forms of abuse, which often co-occur and vary in intensity, undermining the well-being, autonomy, and safety of victims.

In our specific setting, these manifestations are documented through narrative reports generated during legal and support processes related to gender-based violence in northern Mexico. These reports reflect both the severity and the nature of the violence experienced. In this context, the automatic analysis of such textual data has the potential to support early detection, prioritization, and intervention.

However, the automatic analysis of these reports presents several challenges. First, the narratives are linguistically rich and heterogeneous, combining formal descriptions with colloquial expressions and region-specific variations. Second, the classification tasks themselves involve nuanced distinctions: severity levels follow an inherent ordinal structure, while types of violence require multi-label categorization with potentially overlapping classes. These characteristics call for models that are both sensitive to semantic subtleties and capable of handling structured prediction settings.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
alberto.algarra@bbva.com (A. Algarra); s.jimenol@gmail.com (S. Jimeno)

It is within this context that the WomenHelp 2026 shared task [1], *Classification of Gender-Based Violence Reports to Support Threatened Women in Northern Mexico*, is proposed as part of IberLEF Forum [2]. The task focuses on the classification of reports on violence against women through two subtasks: (i) the ordinal classification of violence severity, and (ii) the identification of the types of violence, independently of their severity.

To this end, the organizers provide two datasets. The first consists of 21,257 reports annotated with an ordinal label ranging from 0 (*Mild*) to 3 (*Severe*). The second dataset contains 6,057 reports annotated with one or more of the following categories: *Economic*, *Physical*, *Patrimonial*, *Psychological*, *Sexual*, *Vicarious*, and *N/A*. Each report has been labeled by three human annotators, and the level of agreement among them is provided as a percentage.

Following the competition protocol, both datasets are split into three random partitions with the same proportions: Training (56%), Development (14%), and Test (30%).

2. Related work

Previous editions of IberLEF have included shared tasks addressing harmful content. For instance, the HOMO-MEX task [3], introduced in 2023, focused on polarity detection in messages directed at the LGBTQ+ community in Mexico, and was later extended to 19 Latin American countries in 2025 as HOMO-LAT [4]. Additionally, the DETESTS-DIS shared task [5] addressed the detection of racist content in journalistic articles. More closely related to violence against women, the MiSonGyny shared task [6], part of IberLEF 2025, focused on the detection of misogynistic content in song lyrics.

Across these tasks, transformer-based architectures have become the dominant approach [7]. More recently, with the rise of large language models (LLMs), approaches leveraging models such as Qwen3-14B [8] and LLaMA-3.1-8B through instruction tuning [9] have also been explored. In parallel, ensemble-based methods combining multiple pretrained models have shown strong performance, as demonstrated by Esteban Montesinos-Cánovas et al. [10], and constitute a key source of inspiration for our work. Within this line, Mixture of Experts (MoE) approaches [11] are particularly appealing, as they offer a flexible trade-off between performance and computational efficiency.

Beyond IberLEF, prior work on violence-related text classification spans a wide range of methodologies. Earlier approaches rely on traditional text representations such as TF-IDF combined with classical machine learning algorithms [12], while more recent studies leverage LLMs to extract features from legal texts, for example in Italian [13].

Focusing on ordinal classification, a variety of approaches have been proposed, including the combination of classification models with post-processing regression methods [14], fuzzy rule learning algorithms based on sequential covering strategies [15], and methods that reduce ordinal regression to a set of binary classification tasks [16, 17].

For multi-label classification with soft labels, particularly in settings where annotations reflect annotator agreement, label distribution learning approaches [18] have been shown to be effective. Additionally, methods that explicitly model correlations among labels [19] are especially relevant in scenarios with overlapping categories.

However, many of these approaches rely either on large-scale models or computationally intensive architectures. In contrast, we propose lightweight yet competitive solutions for both subtasks. For the first subtask, we introduce a Mixture of Experts model based on pretrained sentence transformers, combined with a Conditional Ordinal Regression for Neural Networks (CORN) framework to effectively model the ordinal nature of the target labels. For the second subtask, we propose a label-wise architecture built on top of a ModernBERT model pretrained in both English and Spanish [20] without requiring additional fine-tuning, which explicitly captures relationships among labels and is trained using a composite loss function that combines binary cross-entropy with Jensen–Shannon divergence to account for annotator variability.

3. System description

3.1. Subtask 1: system description

For the ordinal severity classification subtask, we propose a training-efficient Mixture of Experts (MoE) architecture that combines representations extracted from multiple pretrained sentence-transformer models.

Figure 1 presents an overview of the proposed Mixture of Experts architecture for ordinal severity classification.

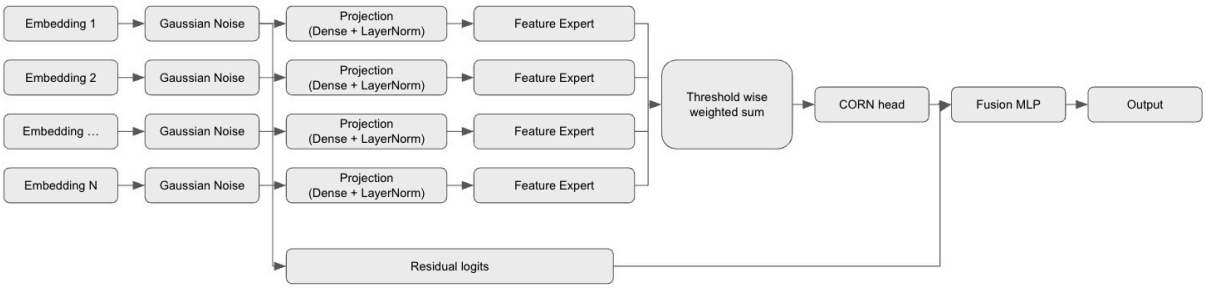

Figure 1: Overview of the proposed MoE architecture. The model combines multiple pretrained embeddings through expert networks and a routing mechanism, followed by CORN-based ordinal prediction and a residual global branch.

3.1.1. Input data

To train the model, we leverage a diverse set of pretrained embedding models, each contributing representations with different dimensionalities.

Model	Dimensionality
intfloat/multilingual-e5-base [21]	768
BAAI/bge-m3 [22]	1,024
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [23]	384
hiiamsid/sentence_similarity_spanish_es [23]	768
sentence-transformers/distiluse-base-multilingual-cased-v2 [23]	512
sentence-transformers/LaBSE [24]	768
sentence-transformers/embeddinggemma-300m-medical	768
sentence-transformers/allenai-specter [25]	768
Tuiaia/bert-base-multilingual-cased-impact-intensity [23]	768
BSC-LT/ModernBERT [20]	768

Table 1
Pretrained embedding models used in our Mixture of Experts architecture and their corresponding feature dimensionality.

This diversity of representations allows the Mixture of Experts architecture to capture complementary semantic signals across domains, languages, and training objectives, improving robustness and overall performance while maintaining a lightweight design.

3.1.2. Internal train/validation split

To construct the training and evaluation splits, from the 13,631 train reports, we adopt a group-aware stratification strategy based on semantic similarity between samples. First, for each pretrained embedding model, we obtain a representation matrix $\mathbf{X}^{(m)} \in \mathbb{R}^{N \times d_m}$, where N is the number of samples

and d_m the embedding dimensionality. Each representation is L2-normalized:

$$\tilde{\mathbf{X}}_i^{(m)} = \frac{\mathbf{X}_i^{(m)}}{\max(\|\mathbf{X}_i^{(m)}\|_2, \epsilon)},$$

where ϵ is a small constant for numerical stability. Then, dimensionality reduction is applied independently to each embedding space using Principal Component Analysis (PCA), yielding $\hat{\mathbf{X}}^{(m)} \in \mathbb{R}^{N \times k}$. The reduced representations are fused by averaging across models:

$$\mathbf{X}_{\text{fused}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{X}}^{(m)},$$

followed by a final L2 normalization step.

To enforce structurally consistent splits, we cluster the fused representations using MiniBatch K-Means. Given $\mathbf{X}_{\text{fused}}$, we obtain cluster assignments:

$$g_i = \arg \min_{c \in \{1,\ldots,K\}} \|\mathbf{x}_i - \boldsymbol{\mu}_c\|_2^2,$$

where $\boldsymbol{\mu}_c$ are the cluster centroids and K is set to $\max(50, \lfloor N/20 \rfloor)$. These cluster labels define groups that are used within a Stratified Group K-Fold splitting procedure, ensuring that samples belonging to the same cluster are not split across folds while preserving the label distribution. In our setting, this results in 680 groups, with an average size of 20.04 samples (median: 17, maximum: 74), providing a balanced trade-off between diversity and structural coherence across splits.

3.1.3. Mixture of Experts

Let $\mathbf{x}_e \in \mathbb{R}^{d_e}$ denote the embedding produced by the e-th encoder. Each embedding is regularized via Gaussian noise and dropout, and projected into a shared space:

$$\mathbf{p}_e = \text{LN}\big(\sigma(\mathbf{W}_e \tilde{\mathbf{x}}_e + \mathbf{b}_e)\big),$$

where $\tilde{\mathbf{x}}_e$ denotes the regularized input. The projected representations are processed by expert networks $f_e(\cdot)$ to obtain feature vectors $\mathbf{h}_e = f_e(\mathbf{p}_e)$. A router takes the concatenated projections and produces label-wise gating weights:

$$g_{e,t} = \frac{\exp(a_{e,t})}{\sum_{j=1}^{E} \exp(a_{j,t})}, \qquad \mathbf{a} = r([\mathbf{p}_1; \ldots; \mathbf{p}_E]),$$

which are used to compute threshold-specific mixtures:

$$\mathbf{m}_t = \sum_{e=1}^{E} g_{e,t} \mathbf{h}_e.$$

The mixture representations are passed through a CORN head to obtain ordinal logits, while a global branch processes the concatenated inputs. Their contributions are combined through a learnable gate $\alpha = \sigma(\alpha_{\text{raw}})$:

$$\hat{\mathbf{y}} = F\big([\text{vec}(\mathbf{M}); \mathbf{q}; \alpha \mathbf{b}]\big), \qquad \hat{\mathbf{y}} \in \mathbb{R}^{K-1}.$$

To prevent expert collapse, we include a load-balancing regularization term:

$$\mathscr{L}_{\text{balance}} = \lambda \sum_{e=1}^{E} \left(\frac{1}{K-1} \sum_{t=1}^{K-1} \mathbb{E}_{\mathscr{B}}[g_{e,t}] - \frac{1}{E} \right)^2.$$

3.1.4. CORN loss

To model the ordinal nature of the severity labels, we employ the Conditional Ordinal Regression for Neural Networks (CORN) loss, which decomposes the ordinal classification problem with K classes into $K-1$ ordered binary classification tasks. Given a ground-truth label $y_i \in \{0, \dots, K-1\}$ and model logits $\mathbf{z}_i \in \mathbb{R}^{K-1}$, each component $z_{i,k}$ represents the logit associated with the conditional probability $\mathbb{P}(y_i > k \mid y_i > k-1)$. For each threshold k, the training set is filtered to include only samples satisfying $y_i > k-1$, and binary labels are defined as:

$$y_i^{(k)} = \mathbb{I}(y_i > k),$$

where $\mathbb{I}(\cdot)$ is the indicator function. The loss is then computed as a weighted binary cross-entropy over all valid thresholds:

$$\mathscr{L}_{\text{CORN}} = \frac{\sum_{k=0}^{K-2} \sum_{i \in \mathcal{S}_k} w_i \cdot \ell_{\text{BCE}}(y_i^{(k)}, z_{i,k})}{\sum_{k=0}^{K-2} \sum_{i \in \mathcal{S}_k} w_i},$$

where $\mathcal{S}_k = \{i \mid y_i > k-1\}$ denotes the subset of samples considered at threshold k, w_i are optional sample weights, and ℓ_{BCE} is the sigmoid binary cross-entropy loss. This formulation ensures that the ordinal structure is respected by enforcing a sequence of conditional decisions, while implicitly encouraging monotonicity in the predicted probabilities across thresholds. The use of sample weights further allows handling class imbalance and varying importance across samples.

3.1.5. Implementation details

The proposed model is implemented in TensorFlow and trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. A learning rate decay strategy is applied, reducing the learning rate by a factor of 0.9 after every five epochs without improvement in the training loss.

All experiments are conducted on a machine equipped with 4 vCPU cores (Intel Xeon 2.20 GHz) and 32 GB of RAM. The proposed Mixture of Experts architecture contains 509,160 trainable parameters and is trained using a batch size of 64, achieving an average training time of 23 ms per step.

The generation of the input embeddings constitutes the most computationally demanding stage of the pipeline, requiring an average of 1.36 seconds per batch. Despite relying on multiple pretrained encoders, the overall approach remains computationally efficient and can be trained using modest hardware resources.

3.2. Subtask 2: system description

For the violence type classification subtask, we propose an architecture built upon embeddings generated by ModernBERT[20], comprising a transformer layer and a combination of global and label-wise attention features, together with label-specific heads and a unified output layer.

Figure 2 presents an overview of the proposed architecture for multilabel classification.

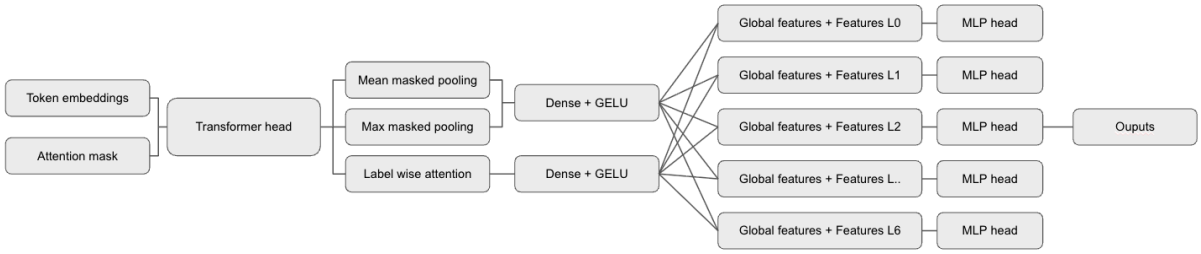

Figure 2: Overview of the proposed Multilabel classifier architecture for types of violence classification.

3.2.1. Input data

For the 4,845 reports in the training set, we extract 768-dimensional embeddings generated by ModernBERT, considering up to a maximum of 533 tokens per report. From these representations, an attention mask is constructed for each sample to prevent the model from learning spurious correlations related to input length rather than semantic content.

The dataset is then split into two subsets (80%/20%) using a stratified sampling strategy adapted to the multi-label setting. Specifically, we employ a multilabel stratified shuffle split method, which extends traditional stratification to multi-label data. Unlike standard approaches that preserve the distribution of a single label, this method iteratively partitions the dataset to maintain the relative frequency of each label across subsets. This ensures that all labels, including rare ones, are proportionally represented in both training and validation sets, leading to a more reliable evaluation in multi-label scenarios.

3.2.2. Label-wise attention

To capture label-specific evidence from the token-level representations, we employ a label-wise attention pooling mechanism. Let $\mathbf{X} \in \mathbb{R}^{T \times d}$ denote the sequence of contextualized token embeddings for a given input, where T is the sequence length and d the hidden dimension. A linear projection is applied to compute label-specific relevance scores:

$$\mathbf{S} = \mathbf{X}\mathbf{W}, \qquad \mathbf{W} \in \mathbb{R}^{d \times L},$$

where L is the number of labels and $\mathbf{S} \in \mathbb{R}^{T \times L}$ contains a score for each token-label pair. When an attention mask $\mathbf{m} \in \{0,1\}^T$ is available, it is used to suppress padded tokens by assigning them a large negative value before normalization:

$$\tilde{S}_{t,l} = \begin{cases} S_{t,l}, & \text{if } m_t = 1, \\ -\infty, & \text{otherwise.} \end{cases}$$

The attention weights are then obtained via a softmax operation over the token dimension for each label:

$$\alpha_{t,l} = \frac{\exp(\tilde{S}_{t,l})}{\sum_{t'=1}^{T} \exp(\tilde{S}_{t',l})}.$$

Finally, a label-specific pooled representation is computed as a weighted sum of token embeddings:

$$\mathbf{h}_l = \sum_{t=1}^{T} \alpha_{t,l} \mathbf{x}_t, \qquad l \in \{1, \dots, L\},$$

yielding a matrix $\mathbf{H} \in \mathbb{R}^{L \times d}$ that encodes distinct contextual summaries for each label. This formulation allows the model to attend to different parts of the input sequence depending on the label of interest, effectively capturing label-dependent semantic cues. In contrast to standard global pooling strategies, which produce a single shared representation for all labels, this approach mitigates information dilution by preserving label-specific relevance patterns. Moreover, by sharing the underlying token representations while learning independent attention distributions, the method remains parameter-efficient and well-suited for multi-label classification scenarios with potentially overlapping classes.

3.2.3. Training objective

For the multi-label classification task with soft labels, we employ a composite loss function that combines a weighted binary cross-entropy (BCE) term with a Jensen–Shannon divergence (JSD) regularization component. Given ground-truth labels $\mathbf{y} \in [0,1]^L$ and predicted probabilities $\hat{\mathbf{y}} \in (0,1)^L$ for L labels, the weighted BCE loss is defined as:

$$\mathscr{L}_{\text{BCE}} = \sum_{l=1}^{L} w_l \left[-y_l \log(\hat{y}_l) - (1 - y_l) \log(1 - \hat{y}_l) \right],$$

where w_l denotes a class-specific weight to account for label imbalance. When logits are used, probabilities are obtained via the sigmoid function, ensuring numerical stability. This term encourages accurate per-label predictions while allowing different importance to be assigned to each class.

To further model the uncertainty and distributional nature of soft labels, we incorporate a Jensen–Shannon divergence term computed independently for each label by treating the predictions as Bernoulli distributions. Let $p_l = (y_l, 1 - y_l)$ and $q_l = (\hat{y}_l, 1 - \hat{y}_l)$, and define the mixture distribution $m_l = \frac{1}{2}(p_l + q_l)$. The JSD for label l is given by:

$$\mathrm{JSD}(p_l \parallel q_l) = \frac{1}{2}\mathrm{KL}(p_l \parallel m_l) + \frac{1}{2}\mathrm{KL}(q_l \parallel m_l),$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence. The final loss is computed as a weighted combination of both components:

$$\mathscr{L} = \alpha\, \mathscr{L}_{\mathrm{BCE}} + \beta \sum_{l=1}^{L} \mathrm{JSD}(p_l \parallel q_l),$$

where α and β control the contribution of each term. This formulation encourages the model not only to fit individual labels accurately but also to align the predicted probability distributions with the soft target distributions, making it particularly suitable in scenarios with annotator disagreement and probabilistic labeling.

3.2.4. Implementation details

The proposed model is implemented in TensorFlow and trained for 30 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3}, which is progressively decayed to 1×10^{-5}.

The experiment is conducted on a machine equipped with 4 vCPU cores (Intel Xeon 2.20 GHz) and 32 GB of RAM. The proposed architecture contains 283,399 trainable parameters and is trained using a batch size of 64, achieving an average training time of 90 ms per step.

4. Results

4.1. Results

Using the folds generated from the training set for each subtask, we compute the same evaluation metrics used in the competition in order to establish a baseline of the system performance before applying it to the blind development and test sets.

Group	Macro F1	Micro F1
KFold	0.6859 ± 0.0251	0.7155 ± 0.0525
Development	0.5410	0.5874

Table 2
Internal cross-validation and official development results for Subtask 1.

Group	Macro F1	Micro F1	Weighted F1	Hamming
KFold	0.7945 ± 0.0612	0.8756 ± 0.0432	0.8745 ± 0.0327	0.0732 ± 0.0074
Development	0.8118	0.8614	0.8646	0.0786

Table 3
Internal cross-validation and official development results for Subtask 2.

For subtask 1, a noticeable drop in performance is observed when comparing the cross-validation folds with the development set, suggesting the presence of a distribution shift (data drift). To further investigate this phenomenon, we perform a visual analysis of the embedding distributions using t-SNE,

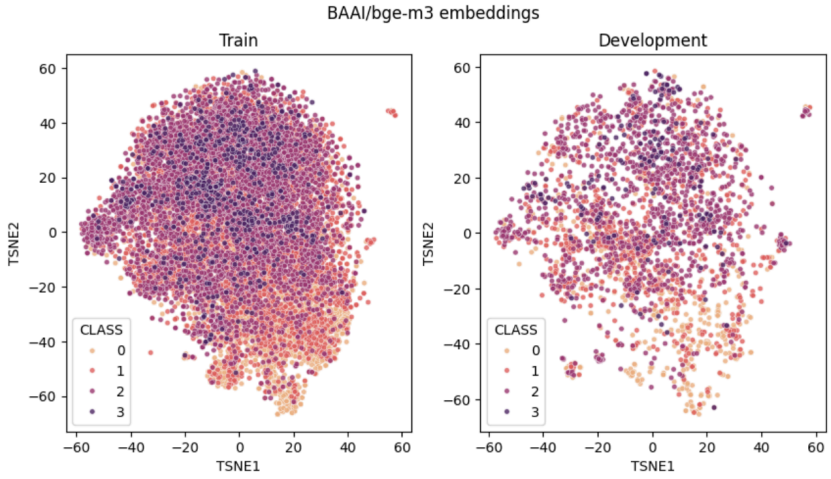

Figure 3: Drift analysis for BAAI/bge-m3 embeddings.

as illustrated in Figure 3. The projection is computed with cosine distance, perplexity 60, random initialization and the automatic learning-rate strategy.

Additionally, we train a simple classifier, namely a logistic regression model, to distinguish between training and development samples. The resulting model achieves an AUC score of 0.86, confirming the presence of data drift. This finding not only explains the observed performance degradation but also suggests a potential direction for future work, such as incorporating reweighting strategies to better align the data distributions.

4.1.1. Subtask 1: Error Analysis

A more detailed analysis of the model behavior is conducted by examining the confusion matrix on one of the training folds, with the aim of understanding the types of errors made by the system:

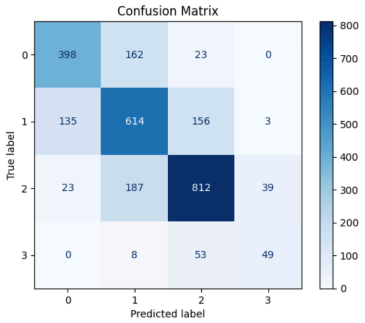

Figure 4: Confusion matrix for a training fold.

Group	Precision	Recall	F1	Support
Development	0.8118	0.8614	0.8646	0.0786

Table 4
Internal cross-validation and official development results for Subtask 2.

As shown in the figure, most of the errors occur between adjacent severity levels, while misclassifications involving jumps of more than one level are rare. This indicates that the proposed architecture effectively captures the ordinal structure of the problem. This behavior is consistent with the use of the CORN loss, which explicitly models ordinal relationships through a sequence of conditional binary decisions.

4.1.2. Subtask 2: Error Analysis

Similarly, confusion matrices are analyzed for each label individually on one of the training folds, with the aim of understanding the model's behavior for each category.

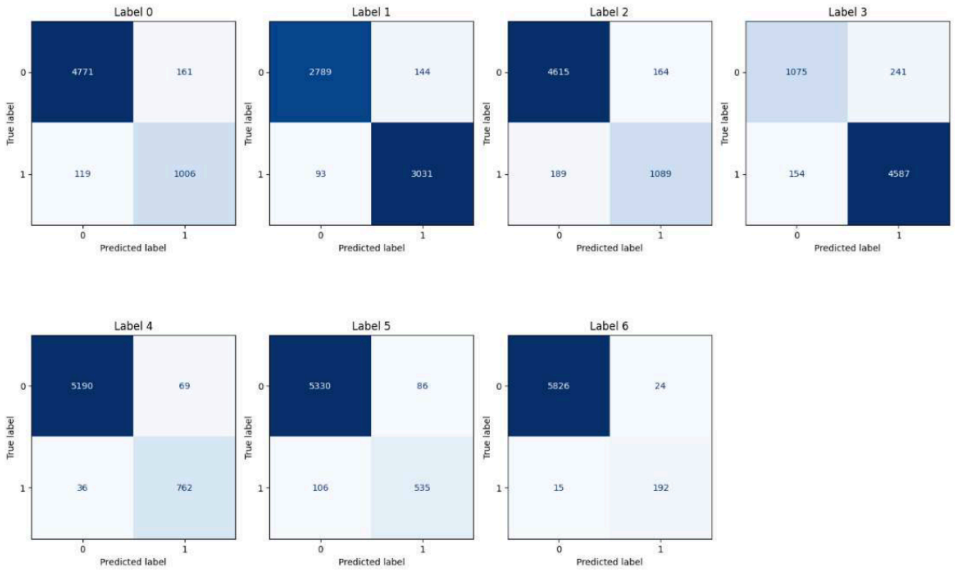

Figure 5: Confusion matrix per label for a training fold.

For a better understanding and comparison, the metrics for each group are presented in Table 5.

Table 5
Metrics per label

Label	Precision	Recall	F1-Score	Positive cases
L0 (Economic)	0.8620	0.8942	0.8778	1125
L1 (Physical)	0.9546	0.9702	0.9624	3124
L2 (Property-related)	0.8691	0.8521	0.8605	1278
L3 (Psychological)	0.9501	0.9675	0.9587	4741
L4 (Sexual)	0.9170	0.9549	0.9355	798
L5 (Vicarious)	0.8615	0.8346	0.8479	641
L6 (N/A)	0.8889	0.9275	0.9078	207

Labels L0 (Economic), L2 (Property-related), and L5 (Vicarious) exhibit the lowest performance. The first two are particularly challenging due to the ambiguity present in many texts, which makes them difficult to identify. In contrast, the performance of the label L3 (Psychological) is mainly due to its high prevalence and the difficulty of correctly identifying negative cases. This highlights the importance of label-specific modeling, as different categories exhibit distinct patterns of ambiguity and imbalance. Future work could focus on refining these categories or exploring modifications to their label-specific heads in order to improve performance.

4.2. Competition leaderboard

The final results on the test set show that our approach remains competitive across both subtasks, particularly in Subtask II, where it achieves performance levels close to those of state-of-the-art systems. Notably, this is accomplished while maintaining a lightweight design, with significantly lower computational requirements and reduced training time. These results highlight the effectiveness of our approach in balancing efficiency and performance, demonstrating that competitive outcomes can be achieved without relying on large-scale or resource-intensive architectures.

Team	S1 MaF1	S1 MiF1	S2 MaF1	S2 MiF1	S2 WF1	S2 Ham
angellabco	0.63	0.62	0.80	0.87	0.87	0.08
franrodrigo	0.62	0.63	0.79	0.87	0.87	0.07
marcelhr	0.61	0.60	0.77	0.85	0.85	0.09
IReL_IIT(BHU)	0.61	0.61	0.78	0.87	0.87	0.08
omargarcia	0.59	0.58	0.78	0.86	0.87	0.08
jorgegomezn	0.59	0.60	0.83	0.89	0.89	0.06
WH_comp_team (ours)	**0.57**	**0.59**	**0.79**	**0.87**	**0.87**	**0.07**
phonghoang	0.57	0.57	0.80	0.87	0.88	0.07
danielrod	0.57	0.57	0.81	0.87	0.87	0.08
manojkumar00011	0.56	0.57	0.78	0.86	0.86	0.08
VerbaNexAI Lab	0.56	0.57	0.82	0.88	0.88	0.07
pln_uam	0.55	0.58	0.81	0.87	0.88	0.07
grupopln	0.54	0.58	–	–	–	–
aerojash	0.57	0.59	0.81	0.87	0.87	0.07
geno_05	0.51	0.55	0.78	0.86	0.86	0.08
UC-UCO-Plenitas	0.48	0.55	0.68	0.77	0.80	0.14

Table 6
Official leaderboard results for both subtasks: S1 MaF1, S1 MiF1, S2 MaF1, S2 MiF1, S2 WF1, and S2 Ham metrics.

5. Conclusion and future work

In this work, we presented two lightweight approaches to address the WomenHelp 2026 shared task. For the ordinal severity classification task, we proposed a Mixture of Experts architecture combined with a CORN loss, effectively modeling the ordered nature of the labels. For the multi-label classification task, we introduced a label-wise attention-based model built on top of ModernBERT embeddings, designed to capture label-specific patterns and handle soft annotations.

Experimental results show that our methods achieve competitive performance across both subtasks while requiring significantly fewer computational resources and shorter training times. Future work will focus on mitigating distribution shift and further improving label-specific components, highlighting the potential of efficient architectures for complex NLP tasks.

6. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-5.3 Instant in order to generate the dot code for graphviz. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards

the mexican spanish speaking lgbt+ population, Procesamiento del Lenguaje Natural 73 (2024) 393–405. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6626.

[4] G. Bel-Enguix, H. Gómez-Adorno, S. Ojeda-Trueba, G. Sierra, J. Barco, E. Lee-Romero, J. Dunstan, R. Manrique, Overview of homo-lat at iberlef 2025: Human-centric polarity detection in online messages oriented to the latin american-speaking lgbtq+ population, Procesamiento del Lenguaje Natural 75 (2025) 413–424. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6764.

[5] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, M. Taulé, Overview of detests-dis at iberlef 2024: Detection and classification of racial stereotypes in spanish - learning with disagreement, Procesamiento del Lenguaje Natural 73 (2024) 323–333. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6620.

[6] T. Alcántara, M. Soto, C. Macias, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveron, Overview of misongyny at iberlef 2025: Misogyny speech detection in spanish language song lyrics, Procesamiento del Lenguaje Natural 75 (2025) 441–451. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6766.

[7] M. Cardoso-Moreno, L. Moreno-Mendieta, D. Jiménez, J. A. Torres-León, J. Valdez-Rodríguez, Homocic at homo-lat 2025: Retrieval augmented classification using transformer architectures and vector knowledgebases, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), Zaragoza, Spain, September 23, 2025., 2025.

[8] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[9] J.-S. Toledo, I. Segura-Bedmar, Hulat_uc3m at misongyny 2025: Detecting misogyny in song lyrics with transformer ensembles and instruction-tuned generative models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), Zaragoza, Spain, September 23, 2025., 2025.

[10] E. Montesinos-Cánovas, F. Garcia-Sánchez, J. A. Garcia-Díaz, G. Alcaraz-Mármol, R. Valencia-García-Sánchez, Spanish hate-speech detection in football, Procesamiento del Lenguaje Natural 71 (2023) 15–27. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6539.

[11] S. Mu, S. Lin, A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications, 2026. URL: https://arxiv.org/abs/2503.07137. arXiv:2503.07137.

[12] E. M. A. Stephanie, L. G. B. Ruiz, M. A. Vila, M. C. Pegalajar, Study of violence against women and its characteristics through the application of text mining techniques, International Journal of Data Science and Analytics 18 (2024) 35–48. URL: https://doi.org/10.1007/s41060-023-00448-y. doi:10.1007/s41060-023-00448-y.

[13] G. Rizzi, D. Scalena, E. Fersini, Gender violence in numbers: Prompting italian llms to characterize crimes against women, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025, pp. 965–973.

[14] E. Frank, M. Hall, A simple approach to ordinal classification, volume 2167, 2001, pp. 145–156. doi:10.1007/3-540-44795-4_13.

[15] J. C. Gámez, D. García, A. González, R. Pérez, Ordinal classification based on the sequential covering strategy, International Journal of Approximate Reasoning 76 (2016) 96–110. URL: https://www.sciencedirect.com/science/article/pii/S0888613X16300706. doi:https://doi.org/10.1016/j.ijar.2016.05.002.

[16] L. Li, H.-t. Lin, Ordinal regression by extended binary classification, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, volume 19, MIT Press, 2006. URL: https://proceedings.neurips.cc/paper_files/paper/2006/file/019f8b946a256d9357eadc5ace2c8678-Paper.pdf.

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

[18] X. Geng, Q. Zhao, Label distribution learning, CoRR abs/1408.6027 (2014). URL: http://arxiv.org/

abs/1408.6027. arXiv:1408.6027.

[19] Y. Zhang, P. Ren, W. Deng, Z. Chen, M. de Rijke, Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3543–3555. URL: https://aclanthology.org/2022.acl-long.248/. doi:10.18653/v1/2022.acl-long.248.

[20] D. Tamayo, I. Lacunza, P. Rivera-Hidalgo, S. Da Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, M. Villegas, Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation, arXiv preprint arXiv:2602.21379 (2026).

[21] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).

[22] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[23] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[24] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, CoRR abs/2007.01852 (2020). URL: https://arxiv.org/abs/2007.01852. arXiv:2007.01852.

[25] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, SPECTER: Document-level Representation Learning using Citation-informed Transformers, in: ACL, 2020.