UMUTeam at WomenHelp 2026: Transformer Models for Severity and Violence Type Classification

Jorge Gómez-Navalón[1], Tomás Bernal-Beltrán[1], Ronghao Pan[1], José Antonio García-Díaz[1,*] and Rafael Valencia-García[1]

[1]*Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain*

Abstract

Violence against women is a complex social problem that manifests through multiple forms of abuse and varying levels of severity. Automatic analysis of textual reports describing such situations can provide valuable support for institutions and professionals involved in victim assistance and decision making. In this paper, we present the UMUTeam submission to the WomenHelp shared task at IberLEF 2026, which addresses two complementary tasks over real anonymized narratives of gender-based violence in northern Mexico: (i) severity classification into four ordinal categories, and (ii) multi-label detection of violence types. Our approach is based on pretrained transformer models fine-tuned for Spanish text classification. For severity classification, we compare three alternative formulations: flat classification, hierarchical classification, and ordinal classification. For violence type detection, we model the problem as a multi-label classification task and optimize the model through hyperparameter search. Experimental results show that standard flat classification achieves the best performance for severity prediction, outperforming more complex ordinal and hierarchical approaches. In the multi-label setting, the proposed model attains strong performance despite class imbalance and overlap among violence categories. The results confirm the effectiveness and robustness of transformer-based models for analyzing real-world narratives of gender-based violence and highlight their potential to support the automatic identification and characterization of high-risk situations.

Keywords

Gender-Based Violence, Violence Against Women, Text Classification, Multi-Label Classification, Ordinal Classification, Transformers, Natural Language Processing, IberLEF 2026

1. Introduction

Violence against women is a pervasive social and public health problem that affects millions of women worldwide and has profound physical, psychological, and socioeconomic consequences. According to the World Health Organization, approximately one in three women globally has experienced physical and/or sexual violence during her lifetime, illustrating the widespread prevalence of the most severe forms of gender-based violence [1]. Beyond physical aggression, gender-based violence encompasses multiple forms of abuse, including psychological, sexual, economic, vicarious, and property-related violence. These manifestations often co-occur and vary considerably in severity, making their identification and characterization a challenging task even for trained professionals [2].

In this context, Natural Language Processing (NLP) techniques offer promising tools for the automatic analysis of textual reports describing situations of violence. Recent advances in transformer-based language models have substantially improved performance across a wide range of text classification tasks, enabling the detection of subtle and context-dependent patterns in natural language [3, 4]. By assisting in the identification of the severity and type of violence described in narratives, computational methods may help support decision-making processes in institutions and organizations that provide assistance to victims [5]. However, this task remains particularly challenging due to the implicit nature

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

jorge.gomezn@um.es (J. Gómez-Navalón); tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

of many descriptions, the overlap between violence categories, and the variability in writing style and linguistic complexity found in real-world reports [6].

The WomenHelp shared task, organized as part of the *IberLEF 2026* evaluation campaign [7], addresses this problem through two complementary classification tasks over anonymized reports of gender-based violence collected in northern Mexico. The first task consists of assigning each report to one of four severity levels (*mild*, *medium*, *high*, and *severe*), while the second task focuses on multi-label identification of the different types of violence present in each report.

In this paper, we present the UMUTeam submission to WomenHelp. Our approach is based on pretrained transformer models fine-tuned for Spanish text classification. For Subtask 1, we investigate three alternative formulations of the severity prediction problem: standard flat classification, hierarchical classification, and ordinal classification. For Subtask 2, we model violence type detection as a multi-label classification task using a transformer-based architecture whose hyperparameters are optimized during model selection.

The experimental evaluation demonstrates that a straightforward flat classification approach outperforms more complex hierarchical and ordinal formulations for severity prediction. In addition, the proposed multi-label model achieves strong performance for violence type detection despite the presence of label imbalance and overlap among categories. These findings highlight the effectiveness and robustness of transformer-based models for analyzing narratives of gender-based violence.

The remainder of this paper is organized as follows. Section 2 reviews previous work related to violence detection and text classification. Section 3 describes the shared task and the dataset. Section 4 presents the proposed methodology. Section 5 reports and discusses the results. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

Gender-based violence against women represents a major social, public health, and human rights problem. In this context, Natural Language Processing techniques have been increasingly explored to support the automatic analysis of large volumes of textual data describing violent situations, including social media content, institutional records, and healthcare narratives. The WomenHelp shared task focuses specifically on the automatic classification of real anonymized narratives describing gender-based violence in northern Mexico.

Previous work has explored the use of NLP for detecting violence-related content in different textual domains. Soldevilla and Flores [8] applied BERT-based fine-tuning to identify gender-based violence messages in Reddit and Twitter, showing that pretrained language models can be effective for detecting violent or abusive content in social media. In a different domain, Botelle et al. [9] evaluated transformer-based NLP models for extracting and classifying references to interpersonal violence from mental healthcare electronic records. Their work is especially relevant because it considers not only the presence of violence, but also additional labels such as victim, perpetrator, domestic violence, physical violence, and sexual violence. This type of formulation is close to the WomenHelp setting, where systems must go beyond binary detection and identify more specific characteristics of the reported situation.

The first subtask of WomenHelp can be understood as an ordinal classification problem, since the labels represent increasing levels of violence severity. This issue has been widely discussed in ordinal classification, where the order among labels is part of the prediction problem. Amigó et al. [10] highlight that common classification metrics often ignore the ordinal structure of the labels, while Cao et al. [11] propose a rank-consistent ordinal regression framework for neural networks. These works motivate the exploration of ordinal-aware strategies for severity prediction, although many practical NLP systems still rely on standard multiclass fine-tuning due to its simplicity and strong empirical performance.

The second subtask is formulated as a multi-label classification problem, because a single report may describe several types of violence simultaneously. Multi-label learning studies scenarios in which each instance can be associated with a set of labels rather than a single class [12]. A common simple strategy

is binary relevance, where one independent classifier is trained per label. However, this approach ignores possible dependencies between labels. Classifier chains were proposed to address this limitation by modelling label correlations through chained predictions [13]. This distinction is important for violence type detection, since categories such as psychological, physical, sexual, economic, property-related, or vicarious violence may appear together in the same narrative and may not be statistically independent.

Transformer-based pretrained language models currently represent the dominant approach for text classification tasks. BERT introduced a general fine-tuning paradigm based on deep bidirectional contextual representations [3], while XLM-R extended this approach to multilingual representation learning at scale [14]. For Spanish, models such as BETO [15], MarIA [16], and RigoBERTa [17] illustrate the benefits of using language-specific pretrained encoders for downstream NLP tasks. These models are especially appropriate for WomenHelp, since the dataset is written in Spanish and includes regional, institutional, and emotionally sensitive language.

A related line of research has focused on sexism, misogyny, and hate speech against women. The AMI shared task at IberEval 2018 introduced automatic misogyny identification in Spanish and English tweets, including misogyny detection, misogynistic behaviour categorization, and target classification [18]. Similarly, HatEval at SemEval-2019 addressed multilingual hate speech detection against immigrants and women in Twitter, with subtasks for binary hate speech detection and finer-grained characterization of aggressive content and targets [19]. More recently, EXIST 2021 proposed sexism identification and sexism categorization in Spanish and English social media texts [20], while EDOS at SemEval-2023 introduced a hierarchical taxonomy for explainable detection of online sexism, covering binary detection, coarse-grained sexism categories, and fine-grained sexism vectors [21]. These benchmarks are relevant because they show the evolution from binary harmful-content detection towards more structured and explainable classification schemes. However, WomenHelp differs from these tasks because it does not focus on public sexist discourse, but on real narratives of gender-based violence reported in institutional contexts.

Overall, previous research shows a clear movement from binary detection of abusive, sexist, or violent content towards more fine-grained and task-specific formulations. WomenHelp follows this direction by combining severity estimation and multi-label violence type detection over real gender-based violence reports. This makes the task more realistic but also more challenging, as systems must handle long narratives, implicit descriptions of abuse, class imbalance, ordinal labels, and possible dependencies between violence categories. Motivated by these challenges, our work adopts a supervised transformer-based approach using Spanish pretrained language models and task-specific prediction strategies for each subtask.

3. Task Description

The WomenHelp shared task, organized as part of IberLEF 2026, focuses on the automatic analysis of gender-based violence reports written in Spanish [22]. The task evaluates the ability of computational systems to identify relevant characteristics of violence narratives, including the severity of the reported situation and the different types of violence present in the text.

3.1. Task Overview

The task is divided into two complementary subtasks:

- **Task 1: Classification by degree of violence.** In this subtask, systems must predict the severity level associated with a violence report. Each instance must be classified into one of four ordinal categories representing increasing levels of severity, ranging from low-risk situations to severe cases requiring urgent attention.
- **Task 2: Types of violence.** This subtask addresses the identification of the different types of violence present in a report. Given a textual narrative, systems must perform multi-label classifi-

cation to determine whether the report contains evidence of categories such as psychological, physical, sexual, economic, property-related, or vicarious violence.

For both subtasks, the input consists of a textual narrative describing a reported situation of gender-based violence. Systems are expected to analyze the linguistic and contextual information contained in the reports in order to produce accurate predictions.

The evaluation of submitted systems is performed using task-specific metrics. Task 1 is evaluated as an ordinal classification problem, where correctly modelling the severity progression between classes is particularly important. Task 2 is evaluated using macro F1-score due to the multi-label nature of the problem and the class imbalance present across violence categories.

3.2. Dataset

The dataset provided by the organizers consists of real narratives describing situations of gender-based violence in northern Mexico. These reports were collected within institutional and social support contexts related to violence against women and were manually anonymized to remove personal or identifying information.

A total of approximately 21k reports are included in the collection, all of them annotated by expert social workers according to the degree of violence described in each narrative. The reports vary considerably in length, writing style, and linguistic complexity, often containing informal expressions, implicit descriptions of abuse, and emotionally charged language.

For Task 1, the complete collection of 21k reports is provided. Each instance is associated with one of four ordinal severity categories representing increasing levels of violence severity: *Mild*, *Medium*, *High*, and *Severe*. This formulation allows systems to model the progression and intensity of violent situations.

For Task 2, the organizers provide a subset of approximately 10k reports annotated for violence type detection. Each instance is labeled using a multi-label scheme covering different forms of violence, including *psychological*, *physical*, *sexual*, *economic*, *property-related*, and *vicarious* violence. Additionally, a *N/A* category is included for reports that do not contain evidence of the predefined violence types. The subset was constructed to preserve the original distribution of severity categories present in the complete dataset.

To illustrate the characteristics of the dataset, Table 1 presents example reports from both subtasks together with their corresponding annotations.

Task 1: Violence Severity Classification
"MENCIONA QUE REALIZO EL TRAMITE DE DIVORCIO Y EL JUEZ FAMILIAR LE ASIGNO UNA PENSION ALIMENTICIA PROVISIONAL Y NO A CUMPLIDO CON EL PAGO ESTABLECIDO."
Severity label: Mild $\rightarrow$ 0
Task 2: Violence Type Detection
"USUARIA REFIERE QUE EL DIA DOMINGO ESTANDO SU VEHICULO ESTACIONADO EN LA ACERA DONDE VIVE LA USUARIA LLEGO EL AGRESOR Y EMPEZO A TIRARLE PIEDRAS AL VEHICULO CAUSANDOLE ABOLLADURAS DEL LADO DEL CONDUCTOR USUARIA ACUDE A MINISTERIO PUBLICO."
Labels: Psychological = 0, Physical = 0, Property-related = 1 Sexual = 0, Economic = 0, Vicarious = 0, N/A = 0

Table 1
Example reports from the WomenHelp dataset for both subtasks.

Both subtasks are divided into training and testing partitions following an 80/20 stratified split. The training partition includes the corresponding annotations, enabling supervised learning, while the test partition requires systems to generate predictions without access to the gold labels. Since the gold labels of the test partition are not available during system development, we further split the labeled

training data into training and development subsets. The development set was used exclusively for hyperparameter optimization, model selection, and validation analysis, while the official test partition was only used to generate the final predictions submitted to the shared task.

The dataset presents several challenges for automatic classification systems, including class imbalance, overlapping violence categories, implicit references to abusive behaviour, and the subjective nature of severity estimation. These characteristics make the task particularly demanding from both linguistic and contextual perspectives.

4. Methodology

This work addresses both subtasks of the WomenHelp shared task using a transformer-based text classification approach built upon pretrained language models for Spanish. The proposed methodology follows a supervised learning pipeline composed of text preprocessing, tokenization, transformer fine-tuning, and task-specific prediction strategies.

For Subtask 1, the problem is formulated as a multi-class severity classification task, where each report is assigned to one of four violence severity levels: *Mild*, *Medium*, *High*, or *Severe*. For Subtask 2, the task is modeled as a multi-label classification problem in which multiple violence categories may co-occur within the same report.

Both subtasks follow a supervised transformer-based learning strategy with task-specific fine-tuning and hyperparameter optimization. Figure 1 illustrates the general workflow of the proposed methodology.

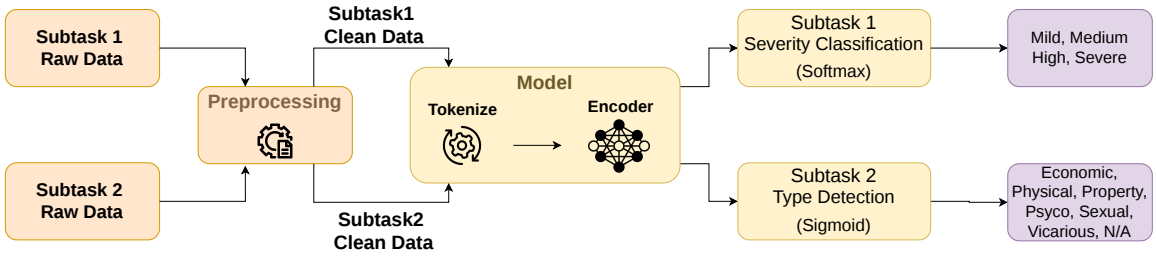

Figure 1: Pipeline of the participation of UMUTeam at WomenHelp 2026.

4.1. Text Preprocessing

Prior to model training, all textual reports undergo a text normalization process designed to reduce noise and improve textual consistency across samples. Given that the dataset contains real-world narratives written in informal language, preprocessing is particularly important to mitigate orthographic variability and formatting inconsistencies.

The preprocessing procedure includes the removal of leading and trailing whitespace characters, normalization of repeated spaces, conversion of all text to lowercase, removal of punctuation symbols, and normalization of quotation marks and apostrophes. These transformations aim to reduce lexical variability while preserving the semantic content of the reports.

After preprocessing, the normalized texts are tokenized using the tokenizer associated with the selected transformer backbone.

4.2. Transformer-based Text Classification

We employ RigoBERTa 2.0 [17] as the backbone model for both subtasks. This model was selected because it is a recent Spanish encoder-based transformer, directly compatible with the HuggingFace fine-tuning pipeline, and suitable for supervised text classification under moderate computational requirements. Since the WomenHelp dataset consists of Spanish narratives containing institutional,

regional, and emotionally sensitive language, we considered a Spanish-specific pretrained model more appropriate than a general multilingual encoder for this setting.

In addition, using the same backbone model for both subtasks allows us to maintain a consistent experimental setup and to focus the comparison on the task-specific formulation of each problem. For Subtask 1, the model is adapted to a multi-class severity classification setting, while for Subtask 2 it is configured for multi-label violence type detection.

The model is fine-tuned independently for each task using the HuggingFace Transformers framework together with a task-specific classification head.

For optimization, we use the AdamW optimizer together with weight decay regularization. Additionally, hyperparameter optimization is performed using Optuna with a Tree-structured Parzen Estimator (TPE) sampler, exploring different values for the learning rate, batch size, number of training epochs, and weight decay. The best configuration is selected according to the macro F1-score on the development set.

4.3. Subtask 1: Severity Classification

Subtask 1 is formulated as a four-class severity classification problem, where each report must be assigned to one of the following categories: Mild, Medium, High, or Severe. Since these labels represent increasing levels of violence severity, we explored three different modeling strategies: flat classification, hierarchical classification, and ordinal classification.

The first strategy, referred to as *flat classification*, treats the task as a standard multi-class text classification problem. In this setting, the model receives the input report and directly predicts one of the four severity categories using a single classification head with four output neurons. A softmax activation is applied over the output logits, and the final prediction corresponds to the class with the highest probability. Although this approach does not explicitly encode the ordinal relationship between the labels, it allows the model to learn direct decision boundaries among all severity levels.

The second strategy, referred to as *hierarchical classification*, decomposes the severity prediction problem into a sequence of simpler decisions. Instead of predicting the four classes in a single step, the model first distinguishes between broader severity groups and then refines the prediction within each group. This formulation is motivated by the assumption that separating low-risk from high-risk situations may be easier than directly distinguishing among all four categories. However, a potential limitation of this approach is error propagation, since an incorrect decision at an upper level of the hierarchy can affect the final severity prediction.

The third strategy, referred to as *ordinal classification*, explicitly considers the ordered nature of the severity labels. In this formulation, the classes are not treated as independent nominal categories, but as levels on an ordered scale. The objective is to encourage the model to distinguish whether a report exceeds different severity thresholds. This approach is intended to penalize distant errors more strongly than adjacent ones and to better reflect the structure of the task, where confusing Mild with Medium is less severe than confusing Mild with Severe.

After preliminary experiments, the standard flat classification approach achieved the best overall performance and was therefore selected as the final system for submission. For this final approach, the reports are tokenized using the RigoBERTa tokenizer with a maximum sequence length of 512 tokens. Sequences longer than this limit are truncated, while shorter sequences are padded to the maximum length.

The system is implemented using the `AutoModelForSequenceClassification` architecture with four output labels. To mitigate the effect of class imbalance, we employ a weighted cross-entropy loss, where the class weights are computed from the label distribution of the training set.

Hyperparameter optimization is performed using Optuna with a TPE sampler. The search process explores different values for the learning rate, batch size, number of epochs, and weight decay. The final configuration uses a learning rate of 2.14×10^{-5}, batch size of 16, 4 training epochs, and a weight decay value of 0.075.

Model selection is performed according to the macro F1-score on the development set. Additionally, accuracy is used as a complementary evaluation metric.

4.4. Subtask 2: Violence Type Classification

Subtask 2 is formulated as a multi-label text classification problem, where each report may contain multiple types of violence simultaneously. Following the task formulation, the system predicts the presence or absence of each violence category independently.

For this subtask, we employ a multi-label classification configuration based on RigoBERTa 2.0. The input reports are tokenized with a maximum sequence length of 256 tokens.

The system is implemented using the `AutoModelForSequenceClassification` architecture configured for multi-label classification. During inference, sigmoid activations are applied to the output logits, and predictions are obtained using a threshold value of 0.5 for each label independently.

Additionally, the system incorporates a simple post-processing strategy for the *N/A* category. Specifically, if no violence category is predicted for a given report, the sample is automatically assigned to the *N/A* class.

Hyperparameter optimization is performed using Optuna with a TPE sampler, exploring different values for the learning rate, batch size, number of epochs, and weight decay. The final configuration uses a learning rate of 2.4275×10^{-5}, batch size of 8, 3 training epochs, and a weight decay value of 0.066.

Model selection is performed according to the macro F1-score on the development set.

5. Results

The experimental results demonstrate the effectiveness of transformer-based approaches for both tasks proposed in WomenHelp. In the first subtask, we compare several alternative formulations of the severity classification problem, allowing us to assess the impact of incorporating ordinal and hierarchical structures into the learning process. In the second subtask, we evaluate a single multi-label classification model and complement the quantitative results with an analysis of prediction errors across violence categories. Together, these experiments provide a comprehensive view of the strengths and limitations of the proposed methodology.

5.1. Subtask 1: Severity Classification

For Subtask 1, we evaluate three different modeling strategies for violence severity classification. All approaches are based on the same pretrained transformer backbone, but differ in the way the classification problem is formulated and optimized.

Approach	Macro-F1 (%)	Micro-F1 (%)
Flat Classification	59.15	59.63
Hierarchical Classification	56.41	58.19
Ordinal Classification	56.47	57.26

Table 2
Results for Subtask 1 severity classification. Metrics are reported as percentages.

The flat classification approach achieved the best performance in both Macro-F1 and Micro-F1 metrics. Although violence severity naturally follows an ordinal structure, explicitly modeling ordinal relationships did not lead to performance improvements over standard multi-class classification.

The hierarchical strategy partially captures the progressive nature of violence severity by decomposing the prediction process into multiple stages. However, the lower performance suggests that error propagation between stages may negatively affect the final predictions.

Figure 2 presents the confusion matrix obtained on the validation set for the best-performing flat classification model. Most classification errors occur between adjacent severity levels, particularly between the *medium* and *high* categories, where a large number of reports are confused in both directions. This behavior indicates that the linguistic distinction between these levels is often subtle and highly context-dependent.

The model also tends to underestimate the most critical situations, as many *severe* cases are predicted as *high*. At the same time, almost no reports are misclassified into distant categories, such as predicting *mild* cases as *severe*. This predominance of near-miss errors confirms that the model captures the ordinal structure of the task despite being trained as a standard multi-class classifier.

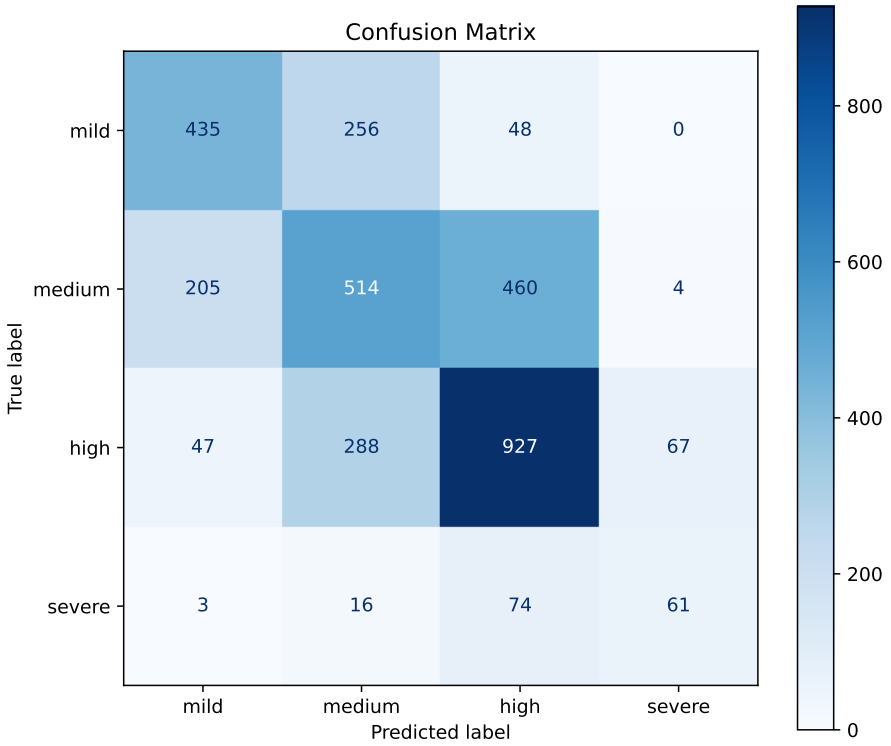

Figure 2: Confusion matrix obtained on the validation set using the best-performing flat classification model for Subtask 1.

Overall, these results indicate that directly modeling the four severity categories is a more robust strategy for this task than introducing additional ordinal or hierarchical constraints.

The system submitted to the official shared task evaluation achieved a Macro-F1 score of 59.15%, ranking 6th out of 16 participating teams. This result is consistent with the validation experiments, confirming that the flat classification strategy generalizes well to unseen data and provides a competitive solution for severity prediction.

5.2. Subtask 2: Violence Type Detection

Metric	Micro-F1 (%)	Macro-F1 (%)	Weighted-F1 (%)	Hamming Loss
Score	88.79	82.68	89.06	0.0631

Table 3
Results for Subtask 2 violence type detection. F1 metrics are reported as percentages, while Hamming Loss is reported on the standard $[0, 1]$ scale.

The proposed transformer-based approach achieved strong performance for the multi-label violence type detection task, obtaining high Micro-F1 and Weighted-F1 scores. These results indicate that the

model is effective in addressing the multi-label classification task despite the presence of label imbalance and category overlap.

The difference between Macro-F1 and Micro-F1 indicates the presence of label imbalance, where frequent categories contribute more strongly to the global performance. This behavior is common in multi-label classification problems involving real-world datasets.

Additionally, the low Hamming Loss value suggests that the model produces relatively few label-wise prediction errors despite the complexity of the task and the possible overlap between violence categories.

Figure 3 presents the number of prediction errors for each violence category. Psychological violence exhibits the highest number of errors, followed by property-related and vicarious violence. In contrast, sexual violence and the N/A category show relatively few errors, suggesting that these labels are associated with more distinctive linguistic patterns and are therefore easier to identify.

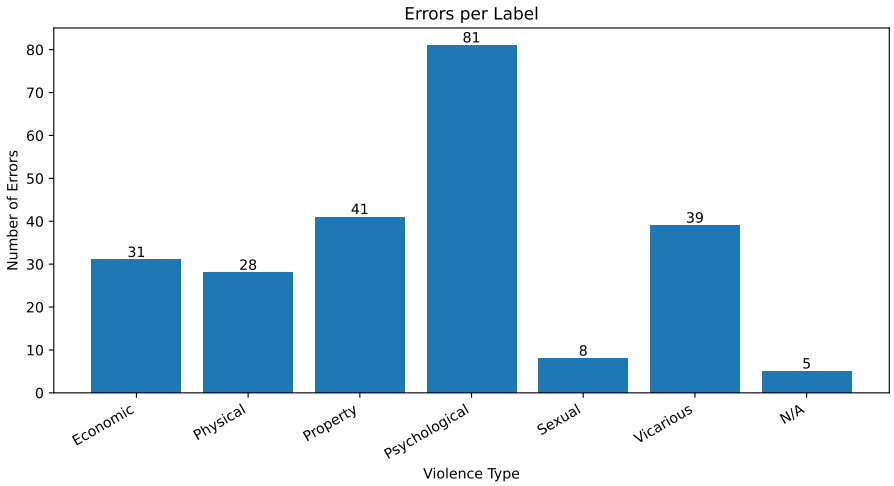

Figure 3: Number of prediction errors for each label in Subtask 2.

Overall, the experiments demonstrate that pretrained transformer models are highly effective for both severity classification and violence type detection in real-world reports of gender-based violence. While standard flat classification proved to be the most robust strategy for ordinal severity prediction, the multi-label formulation achieved strong performance despite class imbalance and label overlap.

In the official shared task evaluation, the submitted system achieved a Micro-F1 score of 88.79% and a Hamming Loss of 0.0631, ranking 1st out of 15 participating teams. This official ranking provides an external comparison with the systems submitted by the other participants and shows that the proposed approach was highly competitive in the shared task setting. Although no additional baseline experiments were included for this subtask, the result suggests that fine-tuning a Spanish pretrained transformer with a multi-label classification head is an effective strategy for identifying co-occurring types of gender-based violence in real-world reports.

6. Conclusion

In this paper, we presented the UMUTeam submission to the WomenHelp shared task at IberLEF 2026, addressing both severity classification and violence type detection in real-world reports of gender-based violence in Mexico. Our approach is based on pretrained transformer models fine-tuned for text classification, with task-specific formulations tailored to the characteristics of each subtask.

The proposed systems achieved highly competitive official results, including a first-place ranking in the violence type detection subtask and a sixth-place ranking in the severity classification subtask.

For Subtask 1, we compared three alternative strategies for modeling violence severity, including flat, hierarchical, and ordinal classification. The experimental results showed that a standard flat classification approach achieved the best performance, outperforming more complex formulations

designed to explicitly capture the ordinal structure of the labels. For Subtask 2, we modeled violence type detection as a multi-label classification problem and obtained strong results, demonstrating that transformer-based models are highly effective despite label imbalance and overlap between violence categories.

Overall, the results confirm the effectiveness and robustness of pretrained transformer models for analyzing narratives of gender-based violence. Beyond their competitive performance, these approaches offer a practical and scalable solution for supporting the automatic analysis of sensitive social data.

Despite these findings, the proposed approach has several limitations. First, the models rely on supervised fine-tuning and therefore depend on the availability, quality, and consistency of expert annotations. Second, the use of fixed maximum sequence lengths may lead to the truncation of relevant information in long narratives, especially when reports describe several events or forms of violence. Third, although most errors in severity prediction occur between adjacent categories, the best-performing model does not explicitly enforce the ordinal structure of the severity labels. Finally, for multi-label violence type detection, the use of a fixed decision threshold for all labels may be suboptimal under strong class imbalance and different label frequencies.

These limitations motivate future work focused on exploring larger language models, ensemble strategies, threshold calibration, and data augmentation techniques to further improve performance, particularly for underrepresented violence categories, long narratives, and challenging intermediate severity levels.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammar and spelling checking.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

[1] World Health Organization, Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women, World Health Organization, 2021.

[2] C. García-Moreno, C. Zimmerman, A. Morris-Gehring, L. Heise, A. Amin, N. Abrahams, O. Montoya, P. Bhate-Deosthali, N. Kilonzo, C. Watts, Addressing violence against women: a call to action, The Lancet 385 (2015) 1685–1695.

[3] 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, volume 1 (long and short papers), 2019, pp. 4171–4186.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017) 5998–6008.

[5] E. Aldana-Bobadilla, A. Molina-Villegas, Y. Montelongo-Padilla, I. Lopez-Arevalo, O. S. Sordia, A language model for misogyny detection in Latin American Spanish driven by multisource feature extraction and transformers, Applied Sciences 11 (2021) 10467.

[6] M. K. D. Schirmer, Natural language processing for violence studies: Investigating trauma and online aggression, Ph.D. thesis, Technische Universität München, 2024.
[7] 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.
[8] I. Soldevilla, N. Flores, Natural language processing through Bert for identifying gender-based violence messages on social media, in: 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), IEEE, 2021, pp. 204–208.
[9] R. Botelle, V. Bhavsar, G. Kadra-Scalzo, A. Mascio, M. V. Williams, A. Roberts, S. Velupillai, R. Stewart, Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study, BMJ open 12 (2022) e052911.
[10] E. Amigó, J. Gonzalo, S. Mizzaro, J. C. de Albornoz, An effectiveness metric for ordinal classification: Formal properties and experimental results, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3938–3949.
[11] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recognition Letters 140 (2020) 325–331.
[12] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering 26 (2013) 1819–1837.
[13] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine learning 85 (2011) 333–359.
[14] 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, 2020, pp. 8440–8451.
[15] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-trained BERT Model and Evaluation Data, arXiv preprint arXiv:2308.02976 (2023).
[16] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, MarIA: Spanish language models, arXiv preprint arXiv:2107.07253 (2021).
[17] I. de Ingeniería del Conocimiento, RigoBERTa-2.0 , 2025. URL: https://huggingface.co/IIC/RigoBERTa-2.0. doi:10.57967/hf/7048.
[18] E. Fersini, P. Rosso, M. Anzovino, et al., Overview of the task on automatic misogyny identification at IberEval 2018., Ibereval@ sepln 2150 (2018) 214–228.
[19] 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, 2019, pp. 54–63.
[20] 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.
[21] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.
[22] 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).