UCO-UC-CICESE-Plenitas Team at WomenHelp 2026: Classification of Gender-Based Violence Reports in Northern Mexico

Yoan Martínez-López[1,2,*], Ireimis de las Mercedes Leguen de Varona[3], Yanaima Jauriga[3], Mayte Guerra Saborit[3], Julio Madera[3], Ana Orquidea López Correoso[4], Luis Quintero Domínguez[5], Ansel Rodríguez-González[6], Kristell Aguilar Alarcón[1,2], Carlos de Castro Lozano[1,2] and Jose Miguel Ramirez Uceda[1,2]

[1]*EATCO. Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*IPVCE "Máximo Gómez Báez", Circunvalación Norte, Camaguey, Cuba*
[5]*Universidad de Sancti Spíritus, Cuba*
[6]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes the participation of the UCO-UC-CICESE-Plenitas team in the WomenHelp 2026 shared task at IberLEF 2026, which addresses the automatic analysis of anonymized narrative reports of gender-based violence collected in the northern region of Mexico. The task comprises two complementary subtasks: severity classification on a four-point ordinal scale (multi-class, single-label) and type-of-violence classification over a closed inventory of modalities (multi-label). We propose a modular hybrid pipeline that represents every report with a single multilingual sentence-embedding backbone and combines three complementary signals: a classical-ML head (a stacking ensemble for the severity subtask and a per-label binary stack with class-wise threshold optimization for the multi-label subtask), an instruction-prompted voting layer over open-weight instruction-tuned LLMs served through Ollama, and a thin layer of lexical heuristics grounded in the Mexican regional register of the corpus. On the official evaluation, our system improved on the shared-task baseline on both subtasks, achieving a macro-F1 of 0.4835 on Subtask 1 (vs. 0.42) and improving over the baseline on Subtask 2. Our internal development and testing analyses identify the multilingual encoder as a second-order but non-negligible factor.

Keywords

Gender-based violence, Text classification, Multi-label classification, Sentence embeddings, Large Language Models, Spanish NLP

1. Introduction

Violence against women is a social problem of profound relevance and global scope, whose manifestations vary according to the historical, cultural and political context of each country and which is, therefore, not exclusive to any single national setting [1]. Far beyond the most visible expressions of physical aggression, gender-based violence encompasses a wide range of modalities —physical, psychological, sexual, economic, vicarious, political or institutional, labour-related and obstetric—

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ymlopez2022@gmail.com (Y. Martínez-López); ireimis.leguen@reduc.edu.cu (I. d. l. M. L. d. Varona); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); mayte.guerra@reduc.edu.cu (M. G. Saborit); julio.madera@reduc.edu.cu (J. Madera); analopezcorreoso@gmail.com (A. O. L. Correoso); laqdominguez@gmail.com (L. Q. Domínguez); ansel@cicese.edu.mx (A. Rodríguez-González); kristell.aguilar@plenitas.com (K. A. Alarcón); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0000-0002-1886-7644 (I. d. l. M. L. d. Varona); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-9556-5869 (M. G. Saborit); 0000-0001-5551-690X (J. Madera); 0000-0002-3527-0516 (L. Q. Domínguez); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-8913-8388 (K. A. Alarcón); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

that can directly or indirectly compromise the physical and emotional integrity of the victim, with degrees of severity that range from low to severe. The breadth of these manifestations underscores the complexity of the phenomenon and reflects how deeply it is entrenched in contemporary societies, where its detection, classification and follow-up remain a pressing challenge for both institutions and research communities. In settings where the volume of incoming reports is high and the resources of the institutions in charge of attending the victims are limited, the automatic analysis of narrative reports of gender-based violence has emerged as a promising line of work [2, 3, 4]. Reports filed at care centres or used as supporting evidence in the justice system tend to be long, lexically rich, written in regional varieties of Spanish, and often produced by social workers and institutional staff rather than by the victims themselves. As a consequence, they exhibit a level of linguistic and stylistic variability that puts them well outside the comfort zone of off-the-shelf sentiment or topic classifiers, while at the same time providing a rich textual signal that, if properly modeled, could support triage, severity assessment and resource allocation in real care workflows.

The WomenHelp 2026 shared task at IberLEF 2026 addresses precisely this scenario[5, 1]. The organizers release a corpus of anonymized narratives drawn from real reports of gender-based violence collected in the northern region of Mexico, in which all personal and identifying information —names of victims, perpetrators and support networks, as well as locations, dates and other potentially identifying details— has been manually removed, while the lexical variation characteristic of the region and the formal style of institutional records are deliberately preserved. The task is framed as a two-subtask supervised classification problem over these narratives [1]:

- Subtask 1 — Severity classification. Each narrative must be assigned a single severity level on a four-point ordinal scale (low, medium, high, severe). The problem is therefore a multi-class single-label classification problem with strong class imbalance, since the most acute severity levels are, by their nature, less frequent in the institutional record than the lower ones.
- Subtask 2 — Type-of-violence classification. Each narrative must be tagged with the modalities of gender-based violence it describes, drawn from a closed inventory that includes physical, psychological, sexual, economic, vicarious, political/institutional, labour-related and obstetric violence. Since a single report can simultaneously document several co-occurring modalities, this subtask is naturally formulated as a multi-label classification problem.

Both subtasks are evaluated with macro-F1, which is the methodologically appropriate choice in this setting: it weights every class —or every label— equally, regardless of its frequency, and thereby prevents a system from obtaining a competitive score by specializing on the dominant categories while ignoring the minority ones, which are precisely the ones that carry the highest social cost when misclassified.

This framing poses several non-trivial research challenges that any participating system must explicitly address. First, the strong class imbalance of the corpus, especially on the severity scale, where the most acute cases are systematically under-represented. Second, the regional and institutional register of the narratives, which combines lexical items typical of northern Mexico with the formal-administrative style of care-process records, and which is markedly different from the general-domain Spanish on which most pretrained encoders have been optimized. Third, the multi-label nature of Subtask 2, which requires the system to capture the co-occurrence patterns of violence modalities rather than treating each label in isolation. Fourth, the sensitivity of the domain itself: the texts describe traumatic events in considerable detail, which conditions both the modeling choices —model outputs must be calibrated and conservative on the high-severity end— and the deployment scenario, in which any system would operate as an assistant to trained personnel rather than as an autonomous decision-maker.

In this paper we present our participation in WomenHelp 2026 with a system designed to address these four challenges jointly. Building on our previous experience in related shared tasks at IberLEF [5], we propose a hybrid pipeline that combines sentence embeddings, augmented by few-shot voting across open-weight Ollama LLMs (qwen, gemma2, deepseek) and a layer of lexical heuristics targeting region-specific markers of violence. The pipeline is trained on the official training partition released by the organizers and evaluated under the official protocol on the held-out test set, with macro-F1 reported separately for each subtask.

2. Methodology

The WomenHelp 2026 shared task is formulated as two complementary supervised classification problems over real, anonymized reports of gender-based violence collected in the northern region of Mexico. Our methodological proposal is built around four design principles that respond directly to the constraints of the task: (i) a single, shared linguistic backbone (a multilingual sentence encoder) is used to represent every report, so that both subtasks consume the same representation and the same pre-processing chain; (ii) the classification head is subtask-specific, with a stacking ensemble for the ordinal multi-class severity problem and a per-label binary stack with class-wise threshold optimization for the multi-label type-of-violence problem; (iii) the classical-ML core is augmented with a few-shot LLM voting layer, which incorporates the world knowledge encoded in open-weight instruction-tuned models without requiring any fine-tuning on the sensitive corpus; and (iv) a thin layer of lexical heuristics, grounded in the Mexican regional register of the corpus, acts as a low-weight but high-precision bias that the ensemble can fall back on when the supervised components are uncertain.

2.1. Subtasks

WomenHelp 2026 is organized into two complementary subtasks, in which each report —a free-text narrative in Mexican Spanish— must be assigned a structured prediction [1]:

- Subtask 1 - Severity classification (multi-class). Each report must be classified into exactly one of four ordinal severity levels: Mild (0), Medium (1), High (2) and Severe (3). The problem is therefore a single-label four-class classification problem with strong class imbalance: the most acute level (Severe) accounts for only $\approx 4\%$ of the training partition, while Medium and High together account for over 70%.
- Subtask 2 - Type of violence (multi-label). Each report must be tagged with the modalities of gender-based violence it documents, drawn from the closed inventory $\{L_0,\ldots,L_6\}$: Economic (L_0), Physical (L_1), Patrimonial (L_2), Psychological (L_3), Sexual (L_4), Vicarious (L_5), and N/A (L_6). Since a single report can simultaneously document several co-occurring modalities, the subtask is formulated as a multi-label classification problem, with the additional constraint that the N/A label is mutually exclusive with the other six —it is reserved for reports that do not explicitly describe any act of violence— and is enforced as a deterministic post-processing rule on the model output.

Internally, both subtasks share the same data ingestion module, the same text-preprocessing chain (whitespace normalization, no lowercasing, no accent stripping, no punctuation removal — the reports are written in upper case with Mexican regional lexicon, and aggressive normalization would discard genuine register signal), and the same embedding backbone.

2.2. Evaluation Protocol

The system outputs are evaluated by the organizers using macro-F1 as the primary measure, computed separately for each subtask. The choice of macro-averaging is methodologically motivated: in both subtasks, the most acute classes —Severe in Subtask 1, and Sexual and violent in Subtask 2— are precisely the minority categories, and a frequency-weighted measure would systematically reward systems that specialize on the majority categories while ignoring the very classes that carry the highest social cost when misclassified. In our internal evaluation, we replicate this protocol on the official development partition released by the organizers and, in addition, report weighted-F1, micro-F1, per-class F1, and Hamming loss (Subtask 2 only) as diagnostic measures. The internal `evaluate_subtask1` / `evaluate_subtask2` routines are implemented on top of sklearn.metrics.`f1_score` (with average="macro", average="weighted", average="micro" and `zero_division=0`) and `sklearn.metrics.classification_report`, so that the in-house numbers are directly comparable with the official leaderboard. For Subtask 2, we additionally report the

F1 of each label individually together with its support on the development partition, which makes it possible to diagnose minority-label failures that would otherwise be hidden inside a macro-average.

2.3. Dataset

The dataset is released by the organisers as two parallel collections of CSV files, one per subtask, distributed under an anonymisation regime in which all personal and identifying information (names of victims, perpetrators and support networks, as well as locations, dates and other potentially identifying details) has been manually removed, while the lexical variation characteristic of northern Mexico and the formal style of institutional records are deliberately preserved.

Subtask 1- A total of 17,038 reports is provided, split into a `training.csv` partition of 13,631 instances and a `devel.csv` partition of 3407 instances. Each row contains a unique `ID`, the report text in column `TEXT`, and the integer severity label in column `CLASS`. The class distribution is strongly imbalanced —approximately 22% Mild, 35% Medium, 39% High and 4% Severe— which directly motivates the use of SMOTE for synthetic minority-class oversampling at training time and of `class_weight="balanced"` in the base learners of the stacking ensemble.

Subtask 2- A total of 6057 reports is provided, split into a `training.csv` partition of 4845 instances and a `test.csv` partition of 1212 instances. Each row contains a unique `ID`, the report text in column `Text`, and seven binary columns $L_0, \ldots, L_6$ that encode the multi-hot label vector. The official labels are the result of an inter-judge aggregation over three annotators, with a label set to 1 whenever its mean across annotators exceeded 0.66 (i.e. at least two of the three annotators agreed); the organizers also distribute a soft version of the labels —the per-document mean annotation scores— as auxiliary supervisory signal, which our pipeline does not currently use but which we discuss as a direction for future work (Section 4). The data ingestion module (`load_subtask1_data` / `load_subtask2_data` in the implementation) reads each CSV with UTF-8 encoding, applies a minimal `preprocess_text` routine that strips leading/trailing whitespace and collapses repeated internal whitespace into a single space, and logs the empirical class (or label) distribution before any model is trained. No further textual normalization is applied.

2.4. Sentence Embeddings

Both subtasks share the same multilingual sentence-embedding backbone within a given experiment, which maps each report into a dense vector representation consumed directly by the downstream classifiers. This shared representation enables a common preprocessing and feature-extraction pipeline across both tasks, while allowing different classification heads to be trained for severity prediction and violence-type detection. During the development phase, multiple multilingual embedding models were evaluated to assess the impact of the backbone on overall system performance.

To represent each report, we explored several multilingual sentence-embedding backbones during the development phase, including BAAI/bge-large [6], sentence-transformers/paraphrase-multilingual-mpnet-base-v2 [7], and intfloat/multilingual-e5-large [8]. All models were implemented through the sentence-transformers library, executed on CUDA when available, and configured to encode inputs in batches of 32 with L2-normalized outputs (`normalize_embeddings=True`). Embeddings were cached in memory on a per-partition basis (train/predict), ensuring that texts were encoded only once during a run and making the per-label binary classifiers of Subtask 2 computationally inexpensive after the initial encoding pass. During development, additional embedding models from the gemma, qwen, nomic and deepSeek families were also evaluated.

2.5. Large Language Models

In addition to the embedding backbone and the classical-ML heads, our pipeline integrates an open-weight instruction-tuned LLM voting layer that operates in few-shot mode. Large Language Models are deep neural networks —overwhelmingly based on the Transformer architecture— trained on massive amounts of text under a self-supervised next-token prediction objective; despite the apparent simplicity

of this training signal, scaling them to billions of parameters and trillions of tokens has produced systems that exhibit broad linguistic competence, multi-step reasoning and the ability to follow natural-language instructions across tasks for which they were not explicitly trained. For WomenHelp 2026, where the corpus is sensitive and the appetite for further fine-tuning is therefore limited, the instruction-following capability of pre-trained LLMs is the property we exploit: each LLM is queried with a structured Spanish prompt that contains the operational definitions taken literally from the official task description, concrete criteria for each severity level (or each violence type), a small number of few-shot examples drawn from the public corpus, and explicit formatting instructions to produce a strict JSON response.

The integration with Ollama is encapsulated in a single OllamaUniversalClient class that uses the /api/chat endpoint (rather than /api/generate) for three reasons: (i) the chat endpoint is supported uniformly by every model in the Ollama catalogue, including gemma [9], qwen [10], and deepseek [11] families; (ii) it allows the format="json" option, which forces the model to emit a syntactically valid JSON object that can be parsed deterministically without ad-hoc regex; and (iii) it correctly applies the model-specific chat templates internally, removing one common source of silent failures across heterogeneous backbones. The client implements exponential-backoff retries (up to three attempts per query), a configurable per-call timeout (120 s by default), and a regex-based JSON parsing fallback for the rare cases in which a model ignores the JSON-mode flag. Decoding is configured with temperature=0.0 and `num_predict=80` (Subtask 1) / `num_predict=120` (Subtask 2), since both subtasks are deterministic classification problems with a single correct answer and creative variation is therefore undesirable.

For the final submission, the LLM voting layer aggregates the predictions of one or more open-weight backbones served locally —the configuration field `ENSEMBLE_LLM_MODELS` admits any list drawn from qwen, gemma and deepseek the three models we benchmarked on the development partition. Each LLM emits a hard label per report, which is converted into a one-hot pseudo-probability vector; the per-model vectors are then averaged across the available LLMs to produce a soft probability matrix that feeds the weighted-voting stage of the ensemble When the LLM voting layer is absent (e.g. when Ollama is unavailable in the execution environment), the corresponding column of the ensemble defaults to a uniform distribution, which makes the system gracefully degrade to an embedding-plus-heuristics configuration without any code change.

2.6. System Architecture

The pipeline is organized as a sequential composition of six decoupled modules, which makes it possible to replace individual components without affecting the rest:

- Module 1 — Data ingestion and preprocessing. Reads the two official CSV partitions for each subtask (training.csv / devel.csv / test.csv), applies the `preprocess_text` routine (whitespace normalization only), logs the empirical class or label distribution, and constructs the typed in-memory dataframes consumed by the rest of the pipeline.
- Module 2 — Embedding backbone. Loads the selected multilingual sentence-embedding backbone onto the available device and exposes an `encode(texts)` method that returns L2-normalized dense vectors in batches of 32, with per-partition caching. In the submitted configuration, the selected backbone is `intfloat/multilingual-e5-large`.
- Module 3 — Classical classification heads. Two subtask-specific heads are built on top of the embedding matrix. Subtask 1 uses an EmbeddingSeverityClassifier that first applies SMOTE over-sampling to the minority severity classes (with `k_neighbors=min(5, min_count-1)` to handle the very small support of the Severe class), and then trains a Stacking-Classifier with three base learners —a calibrated LinearSVC (CalibratedClassifierCV with method="sigmoid", cv=3), an XGBClassifier (400 estimators, `max_depth=6`, `learning_rate=0.05`, `subsample=0.85`, `colsample_bytree=0.85`, `reg_alpha=0.5`, `reg_lambda=1.0`, `min_child_weight=5`), and an LGBMClassifier (500 estimators, 63 leaves, `learning_rate=0.05`, `min_child_samples=20`, `class_weight="balanced"`)— combined through a Logistic Regression meta-learner with 5-fold cross-validation and

`class_weight="balanced"`. Subtask 2 uses an EmbeddingMultiLabelClassifier that trains one LGBMClassifier per label (500 estimators, 63 leaves, `learning_rate=0.05`, `class_weight="balanced"`), with per-class decision threshold optimization in the range [0.20,0.70] at step 0.02 on a 15% internal validation split, maximizing macro-F1 for each label independently. After threshold selection, the binary classifiers are refit on the full training partition before inference.

- Module 4 — few-shot LLM voting layer. Wraps one or more OllamaUniversalClient instances behind a common interface. For each report, every LLM emits a hard label that is converted to a one-hot vector and averaged across the available models. The prompts (`SYSTEM_PROMPT_SEVERITY` / `USER_PROMPT_SEVERITY` for Subtask 1, and `SYSTEM_PROMPT_TYPES` / `USER_PROMPT_TYPES` for Subtask 2) contain the operational definitions of every class taken literally from the official task description, concrete decision criteria for each severity level, few-shot exemplars derived from the public corpus, and explicit instructions to respond in strict JSON.
- Module 5 — Lexical heuristics layer. A thin, low-weight bias built on top of a small set of regular expressions covering the Mexican regional register of the corpus. The lexical patterns were manually designed after inspecting the training corpus and identifying recurrent violence-related expressions in Mexican Spanish:`LEX_PHYSICAL` (physical aggression markers: GOLPEÓ, EMPUJÓ, PATAD, ESTRANGUL, CUCHILLO, ARMA, MORETON…), `LEX_SEXUAL`, `LEX_THREAT_ARM` (death threats with weapons), `LEX_ECONOMIC`, `LEX_PATRIMONIAL`, `LEX_PSYCH`, `LEX_VICARIOUS` (child-related coercion patterns), and `LEX_ADMIN_ONLY` (administrative-only reports that contain no act of violence and that should be biased towards Mild or N/A). The matches are translated into a soft probability vector over the four severity levels (Subtask 1) or the seven labels (Subtask 2), with hand-set scores that have been calibrated on the development partition. This layer is deliberately given a small weight in the final ensemble; it acts as a high-precision back-off when both the embeddings and the LLMs are uncertain, but never overrides them in confident regions.
- Module 6 — Weighted-voting ensemble and submission writer. The three probability streams produced by Modules 3-5 are combined linearly with weights $w_{\text{emb}} = 0.55$, $w_{\text{llm}} = 0.35$, $w_{\text{heur}} = 0.10$ for both subtasks, and the final decision is taken as $\arg\max$ over the resulting probability vector (Subtask 1) or as the elementwise comparison against the class-wise tuned thresholds (Subtask 2), followed by the deterministic mutual-exclusion rule on the N/A label. The two submission files (`subtask1.csv`, `subtask2.csv`) are then serialised in the exact format required by the official CodaBench submission protocol and packaged into a single `submission.zip` archive.

The full configuration —the selection of the embedding model, the list of Ollama backbones, the ensemble weights, SMOTE, the run mode (dev vs. test) and the path layout— is centralized at the head of the entry point, making any experiment fully reproducible from its invocation environment.

Figure 1 summarizes the end-to-end workflow of the pipeline, organized as the six sequential stages.

3. Results

Table 1 reports the internal comparison carried out during the development phase, in which several sentence-embedding backbones were evaluated under the official metric and contrasted against the shared-task baseline. All configurations share the same downstream classifier and differ only in the embedding model used to represent the input, so the table isolates the contribution of the encoder.

Three encoders share the top score of 0.47 —gemma-emb, nomatic-ext and deepseek-emb— followed by a middle cluster at 0.44 (alibaba-qwen, bae-large, gte-qwen) and a lower pair at 0.43 (intfloat-mult, mpnet-base-v2). Every configuration improves on the baseline (0.42), with the three leading encoders exceeding it by 0.05 absolute points (a relative improvement of $\approx 11.9\%$) and even the weakest configurations improving on it by 0.01.

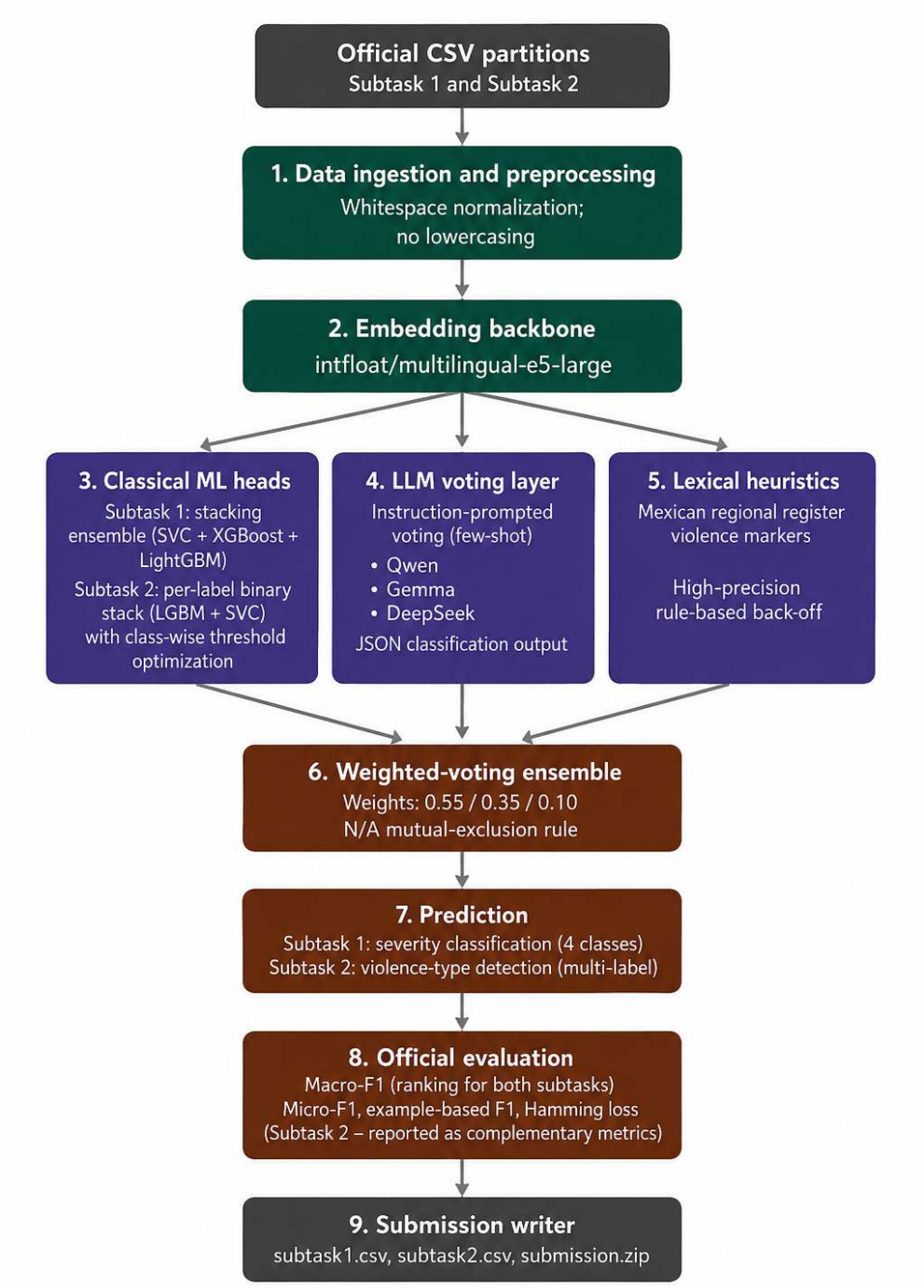

Figure 1: Overview of the proposed WomenHelp pipeline.

Figure 1: Overview of the proposed hybrid pipeline for the WomenHelp 2026 shared task. Reports are encoded using a multilingual sentence-embedding backbone and classified through the combination of classical machine-learning models, instruction-prompted LLM voting, and lexical heuristics.

The most relevant outcome of the development phase is that the embedding-based approach surpasses the official baseline regardless of the encoder chosen: the eight backbones span a narrow range of 0.43-0.47, all at or above the 0.42 baseline. This consistency suggests that the gain comes from the modeling strategy as a whole rather than from any single encoder, and that the choice of backbone is a second-order —though not negligible— factor. The 0.05-point margin of the leading encoders, while

Table 1
Internal Comparisons-Development Phase

Method	Macro-F1
gemma-emb	0.47
nomatic-ext	0.47
deepseek-emb	0.47
intfloat-mult	0.43
alibaba-qwen	0.44
mpnet-base-v2	0.43
bae-large	0.44
gte-qwen	0.44
baseline	0.42

moderate in absolute terms, is obtained without any change to the downstream classifier, which makes it an inexpensive and robust improvement.

The differences across encoders are small but consistent. The leading group (gemma-emb, nomatic-ext, deepseek-emb) reaches 0.47, while the multilingual MPNet (mpnet-base-v2) and the intfloat-mult encoder trail at 0.43. The narrow 0.04-point spread between the best and worst encoders indicates that, for this task, the representation quality of recent multilingual encoders has largely converged, and that switching among them yields diminishing returns. On this basis, and given the three-way tie at the top, the final encoder selection can reasonably be guided by secondary criteria —inference cost, memory footprint, and licensing— rather than by the development score alone.

Table 2 reports the internal comparison carried out during the testing phase, in which the same sentence-embedding backbones evaluated on the development partition were re-evaluated on the test partition and again contrasted against the official baseline. As before, all configurations share the same downstream classifier and differ only in the encoder, so the table isolates the contribution of the embedding model.

Table 2
Internal Comparisons-Testing Phase

Method	Macro-F1
gemma-emb	0.45
nomatic-ext	0.47
deepseek-emb	0.41
intfloat-mult	0.48
alibaba-qwen	0.44
mpnet-base-v2	0.47
bae-large	0.47
gte-qwen	0.47
baseline	0.42

The best test score (0.48) is obtained by intfloat-mult, closely followed by a large group at 0.47 —nomatic-ext, mpnet-base-v2, bae-large and gte-qwen— while gemma-emb (0.45), alibaba-qwen (0.44) and deepseek-emb (0.41) trail behind. Eight of the nine rows improve on the baseline (0.42); the only exception is deepseek-emb, which falls below it by 0.01. The leading encoder exceeds the baseline by 0.06 absolute points (a relative improvement of $\approx 14.3\%$).

The testing phase confirms the main finding of the development phase: the embedding-based approach surpasses the official baseline for almost every encoder, with the best configuration improving on it by 0.06 points. The single configuration that does not (deepseek-emb, 0.41) is also the one that degrades most between phases, which suggests that its development-phase score may have been optimistic rather than that the approach itself is weaker than the baseline.

Table 3 reports the final comparison between our submitted configuration and the official baseline on the test partition across the two subtasks of the shared task. The submitted configuration uses the intfloat-mult encoder, which was retained as the final backbone following the internal development experiments described in the previous section.

Table 3
Comparisons vs baseline-Testing Phase

Method	Subtask 1 Macro-F1	Subtask 2 Micro-F1
intfloat-mult	**0.48**	**0.76885**
baseline	0.42	0.74676

Subtask 1. Our configuration obtains a score of 0.48 against the baseline's 0.42, an absolute improvement of 0.06 points (a relative improvement of $\approx$ 14.3%). This confirms, on the official metric and the test partition, the advantage of the embedding-based approach already observed during the development and testing phases.

Subtask 2. Our configuration obtains a score of 0.7689 against the baseline's 0.7468, an absolute improvement of 0.0221 points (a relative improvement of $\approx$ 3.0%). The improvement is therefore present on both subtasks, although its magnitude differs markedly between them.

Two observations are worth highlighting. First, our system improves on the baseline on both subtasks simultaneously, which is the most desirable outcome and indicates that the gains of the approach are not specific to a single problem formulation. Second, the relative margin is substantially larger on Subtask 1 ($\approx$ 14.3%) than on Subtask 2 ($\approx$ 3.0%). This contrast is consistent with the difference in the absolute difficulty of the two subtasks: the baseline score on Subtask 2 (0.7468) is already high, leaving comparatively little headroom for improvement, whereas the lower baseline on Subtask 1 (0.42) signals a harder problem on which a better representation can contribute a larger margin. In other words, the smaller relative gain on Subtask 2 should not be read as a weakness of the approach, but as a consequence of an already strong baseline near the upper range of the metric.

The official evaluation results released by the organizers are reported for the two subtasks of WomenHelp 2026. Subtask 1 (severity classification) is ranked by macro-F1 and gathered 16 participating teams; Subtask 2 (type-of-violence classification) is ranked by micro-F1, with Hamming loss reported as a secondary measure, and gathered 47 teams/participants. Our team, UCO-UC-CICESE-Plenitas (ymlopez), submitted runs to both subtasks.

Our team, UCO-UC-CICESE-Plenitas, ranked 16th on Subtask 1 (macro-F1 0.4835) and 15th on Subtask 2 (micro-F1 0.7689, Hamming loss 0.1406). On both subtasks the system improved on the official baseline (by $\approx$ 15% and $\approx$ 3% relative, respectively) but remained below a strong and tightly packed field, with the multi-label subtask identified as the component offering the largest margin for improvement.

4. Conclusions

This paper presented the UCO-UC-CICESE-Plenitas submission to the WomenHelp 2026 shared task, which focuses on the automatic analysis of anonymized reports of gender-based violence collected in northern Mexico. Our approach was based on a modular hybrid architecture that combines multilingual sentence embeddings, subtask-specific classification heads, a few-shot LLM voting layer, and a lightweight set of region-oriented lexical heuristics. The same pipeline was applied to both subtasks, providing a unified framework for severity classification and violence-type detection.

The experimental results show that the proposed approach consistently improves upon the official baseline in both subtasks. On the official evaluation, the submitted system achieved a macro-F1 of 0.4835 in Subtask 1 and a micro-F1 of 0.7689 in Subtask 2, corresponding to relative improvements of approximately 15% and 3%, respectively, over the baseline. Internal experiments further indicated that the overall modelling strategy contributes more to performance than the specific sentence-embedding

backbone employed, as most evaluated encoders yielded competitive results within a relatively narrow performance range.

At the same time, the official rankings reveal substantial room for improvement with respect to the leading systems. While the proposed architecture proved effective for severity classification, the multi-label violence-type classification task remains the most challenging component of the pipeline. In particular, the gap in Hamming loss relative to the top-ranked teams suggests that future work should focus on improving label dependency modelling and reducing per-label prediction errors.

Future research will explore stronger multi-label learning strategies, improved threshold calibration, and the incorporation of the soft annotation information released by the organizers. Additional directions include domain-adaptive embedding models and more effective integration mechanisms between embedding-based classifiers and LLM-based reasoning components. Overall, the results demonstrate that a modular hybrid approach constitutes a viable strategy for the automatic analysis of gender-based violence reports, while identifying clear directions for future improvement.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Opus 4.7) and Grammarly for language revision, grammar and spelling correction, and assistance in the preparation of figures.

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. J. Cuevas Hernández, C. A. González González, Emotions and gender-based violence in Mexico, Emotions and Society 7 (2025) 366–382. doi:10.1332/26316897Y2024D000000045.

[3] A. L. S. Hernández, M. A. M. Martínez, F. D. Estrada (Eds.), Gender-Based Violence in Mexico: Narratives, the State and Emancipations, Taylor & Francis, 2023.

[4] K. Ramage, E. Stirling-Cameron, N. E. Ramos, I. Martinez SanRoman, I. Bojorquez, A. Spata, S. M. Goldenberg, "when you leave your country, this is what you're in for": Experiences of structural, legal, and gender-based violence among asylum-seeking women at the Mexico–US border, BMC Public Health 23 (2023) 1699. doi:10.1186/s12889-023-16538-2.

[5] 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 Workshop Proceedings, CEUR-WS.org, 2026.

[6] A. Gupta, M. Rawat, A. Stolcke, R. Pieraccini, REFINE on scarce data: Retrieval enhancement through fine-tuning via model fusion of embedding models, in: Australasian Joint Conference on Artificial Intelligence (AJCAI 2024), Springer Nature Singapore, Singapore, 2024, pp. 73–85. doi:10.1007/978-981-96-0351-0_6.

[7] T. Vrbanec, A. Meštrović, Comparison study of unsupervised paraphrase detection: Deep learning—the key for semantic similarity detection, Expert Systems 40 (2023) e13386. doi:10.1111/exsy.13386.

[8] T. İzdaş, Ö. Sancak, H. T. Kesgin, M. K. Yüce, M. F. Amasyalı, Turkish-E5: E5 model enhanced for Turkish with multi-positive contrastive learning, in: 2025 33rd Signal Processing and Communications Applications Conference (SIU), IEEE, 2025, pp. 1–4. doi:10.1109/SIU66497.2025.11122334.

[9] Gemma Team, Gemma: Open models based on Gemini research and technology, arXiv preprint arXiv:2403.08295 (2024). URL: https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[10] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, others, Z. Qiu, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. doi:10.48550/arXiv.2505.09388.

[11] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Ding, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Chen, J. Yuan, J. Tu, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. You, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Zhou, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, Deepseek-r1 incentivizes reasoning in llms through reinforcement learning, Nature 645 (2025) 633–638. URL: http://dx.doi.org/10.1038/s41586-025-09422-z. doi:10.1038/s41586-025-09422-z.

A. Online Resources

The results of the WomenHelp 2026 are available via:

- WomenHelp,
- WomenHelp 2026 Results,
- WomenHelp 2026 Code