Aerojash at WomenHelp2026: Learning from Disagreement in Spanish Gender-Based Violence Reports

Alix Erica Rojas[1,*], Luis Fernando Niño[1] and Fabio Augusto González[1]

[1]*Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia*

Abstract

Automated classification of gender-based violence (GBV) reports is challenging because severity labels are highly imbalanced and type-of-violence annotations often differ across annotators. This paper describes the *aerojash* submission to the WomenHelp2026 shared task, which addressed severity classification (ST1) and multi-label violence-type classification (ST2) in Spanish GBV reports. Our system combines frozen and fine-tuned multilingual encoders, soft-label training for ST2 using three-annotator label distributions, a severe-class cascade for ST1, and a stacked severity model that uses predicted violence-type probabilities as additional features. The final submission ranked 7th of 16 teams in ST1, achieving a macro-F1 of 0.5750, and 4th of 16 teams in ST2, achieving a micro-F1 of 0.8735.

Beyond the official results, we report observations from system development. The selected soft-label configuration showed its largest per-label gains for the Patrimonial and Vicarious categories, where annotator disagreement was most frequent. In ST1, most errors occurred at the moderate/severe boundary. The reported stacked models achieved competitive development performance but showed sizable development-to-test drops. Although the final ST1 system used ST2 probabilities as additional features, their individual contribution was not evaluated through a dedicated ablation. We present these observations as practical lessons for future shared-task participants and researchers working on socially sensitive natural language processing (NLP) tasks.

Keywords

gender-based violence, Spanish NLP, learning from annotator disagreement, soft labels, class imbalance, ordinal classification, cross-task features

1. Introduction

Gender-based violence (GBV) remains a major social problem, and victim-assistance services often manage large volumes of textual reports describing incidents, threats, and patterns of abuse. Automatic analysis of these reports can support triage, resource allocation, and the identification of violence characteristics. However, the language used in GBV reports is often complex, subjective, and context-dependent, making classification challenging even for human annotators.

The WomenHelp2026 shared task [1, 2] addresses this problem through two related classification tasks over Spanish-language reports collected by victim-assistance offices in northern Mexico. Subtask 1 (ST1) focuses on severity classification, assigning reports to four levels ranging from no violence to severe violence. Subtask 2 (ST2) addresses multi-label classification of violence types, including psychological, physical, economic, patrimonial, sexual, and vicarious violence. The task is particularly interesting because it combines a strong class imbalance in severity prediction with the availability of multiple annotations per instance for violence-type classification.

During system development, two characteristics of the dataset repeatedly influenced model behavior. First, ST2 provides annotations from three annotators, making it possible to observe disagreement patterns that are normally discarded when labels are converted into a single majority-vote target. Second, ST1 is dominated by a difficult boundary between moderate and severe cases, which shaped many of our modeling decisions. Together, these properties informed the design of our final system and the analyses reported in this paper.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

aerojash@unal.edu.co (A. E. Rojas)

0000-0002-0371-3925 (A. E. Rojas); 0000-0003-4703-0007 (L. F. Niño); 0000-0001-9009-7288 (F. A. González)

Participating under the team name *aerojash*, we explored a range of approaches, including frozen multilingual embeddings, transformer fine-tuning, imbalance-handling strategies, severity-oriented cascades, and stacked models. The final submission ranked 7th of 16 teams in ST1 and 4th of 16 teams in ST2. Beyond the official rankings, the development process provided an opportunity to examine several practical questions that frequently arise in shared-task settings: how to make use of annotator disagreement, how severity imbalance affects model behavior, and how stacked ensembles translate from development data to official test evaluation.

This paper has two goals. First, we describe the system submitted to WomenHelp2026 and the design decisions that led to the final results. Second, we report a set of empirical observations and lessons learned during development. In particular, we analyze the use of soft-label training based on annotator disagreement, discuss the incorporation of violence-type predictions as severity cues within the selected system, and examine the reliability implications of imbalance handling, stacking, and model selection. Rather than presenting these findings as definitive conclusions, we report them as observations from a competitive shared-task setting that may be useful to future participants and researchers working on socially sensitive NLP classification tasks.

The remainder of the paper is organized as follows. Section 2 reviews related work, and Section 3 introduces the task and dataset. Section 4 describes the system design, modeling decisions, and experimental setup. Section 5 reports the main results, while Section 6 discusses the lessons learned during system development. Finally, Section 7 concludes the paper.

2. Related Work

Research on gender-based violence (GBV) and related forms of abusive language has attracted growing attention within NLP. A substantial body of work has addressed the automatic identification of abuse, harassment, hate speech, and other violence-related content across social media, online platforms, and institutional records; the detection of hate speech targeting women, in particular, has motivated dedicated multilingual benchmarks in Spanish and English [3]. More recently, this line of research has extended toward supporting victim-assistance services and analyzing gender-based violence reports directly, where the central difficulties are less lexical than interpretive: the narratives are subjective, meaning is strongly context-dependent, and the categories of interest are severely imbalanced. The WomenHelp task situates itself within this trajectory, shifting the focus from social-media abuse to severity assessment and violence-type classification in Spanish-language reports drawn from real victim-assistance settings.

A recurring challenge in socially sensitive classification tasks is annotator disagreement. When labels depend on subjective judgments or complex social phenomena, different annotators may interpret the same text differently while reaching equally valid conclusions. A growing body of work argues that such disagreement should not be treated as annotation noise, since minority judgments often capture alternative but meaningful interpretations of the data [4, 5]. This perspective has motivated methods that move beyond majority voting by preserving the full label distribution or training directly on soft targets derived from multiple annotations [6].

The WomenHelp dataset makes this perspective especially relevant. Each report receives three independent annotations for violence-type classification, so labels assigned by only one annotator are discarded by majority voting. Rather than treating these cases as annotation errors, we view them as additional supervision for categories with ambiguous boundaries. This perspective directly motivates our use of soft labels.

Class imbalance represents a second, largely orthogonal challenge in both GBV classification and NLP more broadly. The minority classes are frequently the most consequential from an operational perspective, yet they are the least represented in the training data. In severity classification, the problem is compounded by ordinal structure: errors do not scatter uniformly across classes but concentrate on adjacent categories. A range of remedies has been proposed, including loss re-weighting, resampling, and specialized objectives such as focal loss [7], although their effectiveness is highly task-dependent.

Finally, a growing body of work exploits information shared across related tasks. Multi-task learning, transfer learning, and the reuse of one model's predictions as inputs to another have all been used when different label sets capture complementary aspects of the same underlying phenomenon. Stacked generalization, in particular, combines the predictions of multiple base models through a meta-learner [8]. In WomenHelp, severity and violence type are plausibly related, since the combination of violence types described in a report may provide information about severity. This relationship motivates the use of violence-type probabilities as one component of our final severity model. Rather than proposing a new learning paradigm, we examine how this design choice, together with disagreement-aware supervision and class imbalance, behaves in a competitive shared-task setting.

3. Task and Dataset Overview

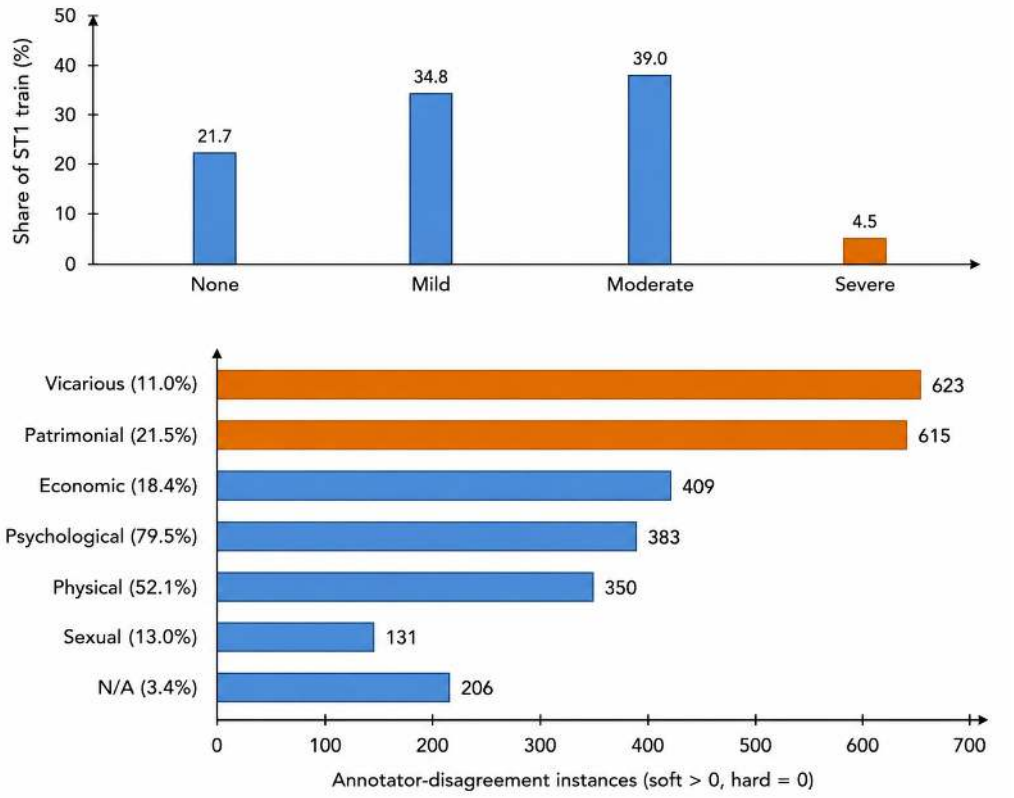

Figure 1: Dataset and annotation profile. The upper chart shows the class distribution for ST1 severity classification in the training set, highlighting the strong imbalance, with the Severe class representing only 4.5% of the instances. The lower chart reports minority-positive disagreement counts for each ST2 violence label (instances with soft > 0 and hard = 0), with overall label prevalence shown in parentheses beside each label and the corresponding number of disagreement instances shown at the end of each bar.

Task definition. Subtask 1 is a four-class single-label severity problem (0: none; 1: mild; 2: moderate; 3: severe), evaluated using macro-F_1. Subtask 2 is a seven-label multi-label type-of-violence problem (Economic, Physical, Patrimonial, Psychological, Sexual, Vicarious, and N/A); its official ranking metric is micro-F_1, and we use macro-F_1 for development-set analysis where per-label behavior is of interest. ST1 comprises 13,630 training, 3,405 development, and 4,219 test reports; ST2 comprises 4,845, 1,212, and 3,620 instances, respectively.

Annotation structure and soft labels. Each ST2 instance is annotated by three annotators, so for every label the fraction of annotators who assigned it takes one of four values: 0, $1/3$, $2/3$, or 1. These soft labels are available for the training and development splits. We define an *annotator-disagreement*

instance for a given label as one in which at least one annotator assigned the label but the majority did not (soft > 0, hard $= 0$); such instances are evidence of disagreement on ambiguous categories [4, 5], not recovered ground truth, and we use them as a supervision rather than as a label-correction oracle. Figure 1 provides an overview of the two dataset characteristics that motivate our approach: the strong class imbalance in ST1 and the distribution of minority-positive disagreement instances across the ST2 violence labels. The severe class is both rare and adjacent to a much larger class, while disagreement is concentrated in Vicarious and Patrimonial.

4. System Design and Modeling Decisions

During system development, two properties of the dataset repeatedly influenced our modeling decisions. First, the availability of three annotations per ST2 instance made it possible to preserve disagreement information instead of collapsing labels into a majority vote. Second, a separate auxiliary ST2 classifier could generate violence-type probabilities as additional features for severity classification. The submitted ST2 ensemble, therefore, uses per-label annotator proportions as continuous training targets, while the submitted ST1 system combines the auxiliary violence-type probabilities with a severe-class cascade and other model outputs together with lexical features in a stacked predictor. We additionally examine development-to-test differences and the severity-confusion structure of the strongest standalone severity model. Figure 2 summarizes these components and the point at which the auxiliary ST2 probabilities enter the severity pipeline.

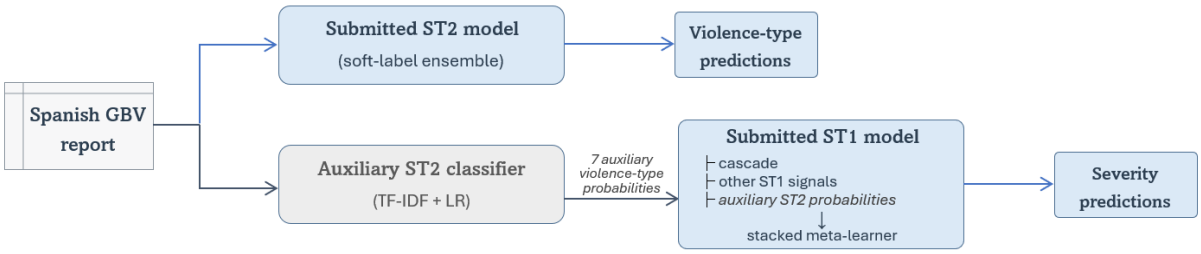

Figure 2: System pipeline. Each Spanish GBV report is processed by the submitted soft-label ST2 ensemble, which produces the official violence-type predictions, and by a separate auxiliary TF-IDF logistic-regression classifier, which produces seven violence-type probabilities for ST1. The submitted ST1 predictor combines these auxiliary probabilities with the severe-class cascade and other ST1 signals through a stacked meta-learner.

4.1. Soft-label learning for violence-type classification

Majority voting maps the three annotations of each instance to a single binary target per label, discarding the information carried by minority annotators. Instead, we preserve the degree of annotator agreement during training. If one, two, or three annotators assign a label, the target reflects that proportion of agreement ($1/3$, $2/3$, or 1) rather than a hard majority-vote decision. The soft targets affect training only; evaluation uses the official hard labels. This treatment does not assume that minority annotators are correct. It treats disagreement as informative uncertainty: where annotators systematically diverge, the model is encouraged to represent that divergence rather than to commit prematurely to the majority decision.

We instantiate this idea in two ways. In the frozen-embedding setting, the encoder parameters remain fixed; mean-pooled sentence vectors are extracted once and used to train an XGBoost regressor with a logistic objective on the soft targets. In the fine-tuning setting, XLM-RoBERTa-large is trained with a binary cross-entropy loss on the seven soft targets. Per-label decision thresholds are averaged over a five-fold cross-validation of the development set rather than tuned on the full split.

4.2. Boundary-aware severity classification

The dominant severity error is structured rather than diffuse. Across the ST1 models evaluated during development, the majority of the severe class's false negatives were predicted as *moderate*, so the severity problem is largely a single adjacent-boundary problem. Standard cross-entropy penalizes a $3 \rightarrow 2$ error identically to a $3 \rightarrow 0$ error, even though the former is both more frequent and, in an ordinal sense, less costly. We therefore decouple the rare-class decision from the rest with a two-stage cascade (Figure 3). The first stage is a dedicated severe-versus-rest detector that emits a probability compared against a learned threshold; reports that clear the threshold are labeled severe, and the remainder are passed to a second stage that distinguishes no violence, mild, and moderate. Separating the rare-class decision lets the system specialize on the severe boundary before making finer distinctions among the non-severe classes. We treat the cascade as a modeling response to ordinal imbalance, not as the central contribution of the paper; it is the strongest standalone severity system, but remains boundary-limited. Both cascade stages are XGBoost classifiers trained on mean-pooled sentence embeddings from a frozen XLM-RoBERTa-large encoder. Stage 1 estimates a severe-versus-rest probability that is compared against a threshold tuned on the development set; Stage 2 predicts among classes 0, 1, and 2 over the reports that fall below the threshold. The cascade outputs serve as one of the base learners feeding the final ST1 stack described in Section 4.5 and are not the submitted predictor.

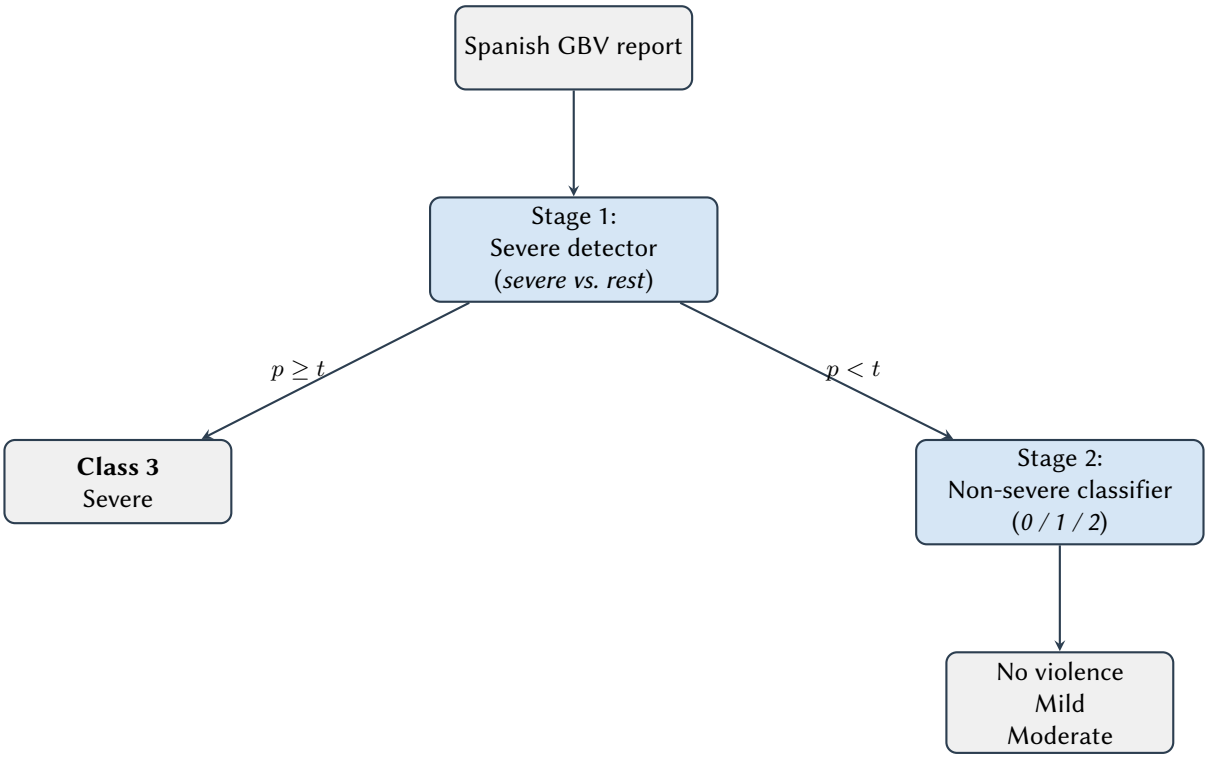

Figure 3: Severe-class cascade for ST1. The system first detects whether a report is severe; only non-severe reports are passed to a second classifier that predicts classes 0, 1, or 2. The design follows directly from the observation that most severe false negatives are confused with moderate.

4.3. Violence-type probabilities as severity cues

One of the design choices explored in the final severity model was the inclusion of cross-task features. For this purpose, we trained a separate auxiliary ST2 classifier using word- and character-level TF-IDF features with one-vs-rest logistic regression. For each ST1 report, this classifier produces seven probabilities, one per violence type. These probabilities are concatenated with the cascade outputs, the outputs of the other ST1 base models, and the lexical features in the input vector consumed by the ST1

meta-learners. The motivating hypothesis is that the composition of violence types carries severity-relevant information—for instance, the co-occurrence of economic and psychological violence may shift the moderate/severe boundary.

We are deliberate about how strongly we claim this. The present experiments do not isolate the marginal contribution of the cross-task features: no leave-one-group-out ablation was run, and the features enter alongside many other signals in a stacked model. We therefore treat the auxiliary violence-type probabilities as a design element of the selected severity model, not as an independently established causal factor.

4.4. Experimental Design

The experimental design was organized around the main observations that emerged during system development:

- **Violence-type classification (ST2).** We compared hard-label and soft-label training using both frozen sentence embeddings and fine-tuned transformer models.
- **Severity classification (ST1).** We compared systems of increasing complexity, ranging from frozen-embedding baselines and the severe-class cascade to ensemble and stacked variants.
- **Evaluation metrics.** Severity classification was evaluated using macro-F_1, while type classification was ranked using the official micro-F_1 metric. We additionally report development macro-F_1 for per-label analysis.
- **Model selection.** Base systems were selected using development-set performance and evaluated on the official test set. For ST1, candidate meta-learners were evaluated using five-fold out-of-fold predictions within the development split; the final weighted vote and class-3 calibration threshold were then selected using development-set performance. For ST2, per-label thresholds were estimated through five-fold cross-validation on the development split.
- **Reliability assessment.** We report development-to-test differences for the evaluated ST1 systems and analyze the severe/moderate confusion pattern. Where controlled comparisons were available, paired-bootstrap resampling was used on the development set.

Development-to-test differences were interpreted cautiously because the final ST1 vote and calibration threshold were selected using development-set performance. Throughout the paper, reliability diagnostics are used to complement aggregate performance scores. All experiments were run on a single laptop; transformer fine-tuning used GPU mixed-precision training.

4.5. Implementation summary

For reproducibility, we briefly summarize the main implementation choices:

- We evaluated three multilingual encoders—BETO [9], XLM-RoBERTa-large [10], and multilingual-E5-large [11]—as frozen sentence-embedding extractors feeding a gradient-boosted baseline [12]. XLM-RoBERTa-large was also evaluated through task-specific fine-tuning.
- The frozen-embedding baseline follows the gradient-boosting configuration released by the task organizers, with class probabilities averaged over five random seeds.
- For ST2, models were trained on soft targets derived from the proportion of annotators assigning each label. Decision thresholds were estimated through five-fold cross-validation on the development set. The submitted ST2 predictor is a four-component ensemble comprising the fine-tuned XLM-RoBERTa-large model reported in the Soft column of Table 1, a 0.5/0.5 blend of that model with the frozen-embedding XGBoost predictor, and two continued-training checkpoints obtained at progressively lower learning rates. This ensemble produces the official violence-type predictions and is distinct from the auxiliary TF-IDF logistic-regression classifier used to generate the seven cross-task probabilities for ST1.

- For ST1, the submitted predictor is a weighted vote over the nine highest-scoring candidate predictors selected using development-set performance. The candidate pool included regularized logistic regression, Extra-Trees, random forest, XGBoost, and simple probability-average variants. The input matrix to these meta-learners concatenates four groups of features: (i) the probabilities produced by the severe-class cascade; (ii) outputs from the other fine-tuned and frozen-embedding ST1 models; (iii) predictions from TF-IDF logistic-regression baselines together with compact lexical and length features; and (iv) the seven violence-type probabilities produced by the auxiliary ST2 classifier. The final weighted vote is followed by a class-3 calibration rule: reports whose aggregated severe-class probability satisfies $(p_3 \geq 0.47)$ are assigned to the severe class. The threshold was selected on the development set.

5. Results

5.1. Largest gains occur on disagreement-heavy labels

On development, the selected soft-label configuration outperformed the hard-label frozen-embedding baseline. Its largest per-label gains were observed for Patrimonial and Vicarious, the two categories with the highest recorded levels of annotator disagreement in the training data, while the high-agreement Psychological category changed only marginally (Table 1).

Table 1
ST2 per-label development F1. *Hard* is the hard-label frozen-embedding (three-encoder) baseline and *Soft* the soft-label fine-tuned model; the two columns therefore differ in both the training target and the model, so these gains reflect the full configuration rather than soft labels in isolation (the controlled, fixed-encoder soft-vs-hard comparison is reported in the text). The largest gains fall on the disagreement-heavy *Patrimonial* and *Vicarious* categories, while the high-agreement *Psychological* category changes only marginally.

Label	Hard	Soft	Gain
Patrimonial	0.645	0.770	$+\mathbf{0.125}$
Vicarious	0.538	0.627	$+\mathbf{0.089}$
Physical	0.865	0.936	$+0.071$
Economic	0.761	0.805	$+0.044$
Sexual	0.783	0.820	$+0.037$
N/A	0.821	0.842	$+0.021$
Psychological	0.911	0.918	$+0.007$
Macro-F_1	0.7605	0.8166	$+0.056$

The per-label gains also reflect each label's baseline headroom: high-prevalence labels such as *Psychological* already score highly and thus have little room to improve. A controlled comparison that holds the encoder fixed isolates the effect of the soft targets themselves: in the fixed-encoder frozen-embedding comparisons, the gain is largest for BETO (+0.135 macro-F1) and shrinks to +0.006 for the concatenation of the three encoders. On the official test set, the type-classification system achieved a micro-F1 of 0.8735, ranking fourth among sixteen participating teams. Together, these results support the value of preserving disagreement during training rather than collapsing annotations to a majority vote—while not implying that minority annotations are correct or that the system is ready for deployment.

5.2. Severity classification remains boundary-limited

The severe-class cascade is the strongest standalone severity system in our experiments, yet its residual errors confirm that severity classification remains largely a boundary problem. Figure 4 shows the cascade's development confusion matrix. Even after dedicated severe-class handling, 58.4% of true-severe reports are predicted as moderate (the highlighted cell): the severe-to-moderate confusion

alone accounts for 90 of the 154 severe instances. The other ordinal neighbors behave similarly, with most mass concentrated on or adjacent to the diagonal. The matrix therefore confirms that the moderate/severe boundary remains the cascade's main source of error and that any intervention targeting the severe class would need to be evaluated against the full per-class profile.

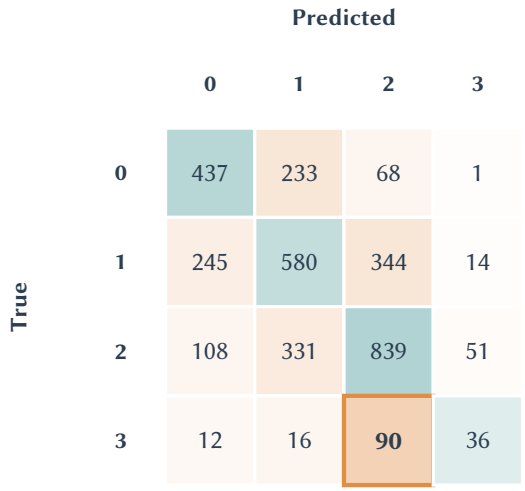

Figure 4: Confusion matrix of the standalone severe-class cascade on the development set; rows are true classes, columns are predicted, and cell shading is row-normalized. This figure analyzes the cascade base learner separately and does not represent the final submitted ST1 stack. The dominant residual error is severe reports predicted as moderate (highlighted): 90 of 154 severe reports.

5.3. Reliability, cross-task features, and model selection

Development and official test performance differed across the evaluated severity systems. The selected stack scored 0.6172 on development and 0.5750 on the official test set, a development-to-test decrease of 0.0422. An alternative stack with a competitive development score of 0.5996 achieved only 0.56 on the official test set. These results show that development performance overstated official-test performance for both reported stacks. For one intermediate ST1 system—the cascade augmented with three auxiliary one-epoch severe-versus-rest detectors—a paired bootstrap comparison against the plain cascade yielded a mean macro-F1 difference of $+0.0075$, with the ensemble outperforming the cascade in 96.7% of bootstrap samples. No comparable bootstrap estimate was available for the stacked systems.

The selected severity model incorporated the seven violence-type probabilities produced by the auxiliary ST2 classifier as additional features. This configuration achieved the best official severity result among the evaluated systems (macro-F1 = 0.5750, 7th of 16) and reached 0.6172 on development. Because the cross-task features were introduced together with several other components in a stacked model, their marginal contribution was not isolated through a dedicated ablation. The result is therefore best interpreted as consistent with the usefulness of cross-task features rather than as direct evidence of its independent effect.

6. Discussion and Lessons Learned

The WomenHelp shared task provided an opportunity to examine how different forms of supervision and model design choices behave in a realistic GBV classification setting. Beyond the official rankings, several consistent patterns emerged during system development and evaluation.

One finding concerns annotator disagreement. The selected soft-label configuration showed its largest gains for Patrimonial and Vicarious, the two labels with the highest recorded levels of annotator disagreement in the training data, whereas the high-agreement Psychological label changed very little. We do not interpret this as evidence that minority annotations are correct. Instead, it suggests that

disagreement captures ambiguity that can be preserved during training, making soft-label learning particularly useful for categories with less clearly defined boundaries.

A second finding is that severity classification is not only an imbalance problem but also a boundary problem. Most residual errors remained concentrated on the moderate/severe boundary rather than being distributed across classes. The confusion matrix in Figure 4 suggests that methods targeting the severe class should be evaluated against the complete per-class error profile, rather than severe-class performance alone.

A third finding concerns reliability. The reported stacked models showed sizable development-to-test drops, indicating that development scores can overstate expected official-test performance. Reporting these gaps, and where possible using out-of-fold meta-features, may provide a more realistic assessment of system behavior than development scores alone.

The use of violence-type probabilities as additional features in the severity model is also noteworthy. The selected system achieved the best official severity result among the evaluated models. However, because these features were introduced together with several other components in a stacked architecture, their individual contribution was not evaluated through a dedicated ablation. We therefore view this result as an encouraging observation rather than evidence of an independent effect.

These observations lead to three practical lessons. First, preserving annotator disagreement can improve performance when category boundaries are ambiguous. Second, methods targeting rare classes should be evaluated across the complete per-class error profile, including adjacent categories. Third, the observed development-to-test drops highlighted the limitations of relying exclusively on development scores.

Ethical and practical considerations. GBV triage is socially consequential, and the systems studied here are not suitable as autonomous decision-makers. The severity model remains fragile on the high-stakes severe boundary, and the gains observed from disagreement-aware training reflect improved representation of ambiguity rather than the resolution of that ambiguity. Such systems should therefore support, rather than replace, the judgment of trained human caseworkers. Any real-world deployment would require careful auditing of error costs, particularly false negatives in severe cases, as well as attention to the institutional and geographic specificity of the data.

7. Conclusion

This paper described the *aerojash* submission to the WomenHelp2026 shared task for severity classification and violence-type classification in Spanish GBV reports. The final system ranked 7th of 16 teams in severity classification and 4th of 16 teams in violence-type classification.

Our experiments led to three main findings. First, the selected soft-label configuration showed its largest gains for the violence categories with the highest levels of annotator disagreement. Second, the severity configuration that incorporated violence-type probabilities achieved the strongest official result among the evaluated systems, although the individual contribution of these features was not isolated. Third, severity classification remained dominated by the moderate/severe boundary, and the observed development-to-test drops highlighted the limitations of relying exclusively on development scores.

We hope that both the system description and the lessons learned are useful to future WomenHelp participants and to researchers working on socially sensitive NLP classification tasks.

Acknowledgments

The authors thank the organizers of the WomenHelp 2026 shared task for providing the dataset.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to support language editing, improve the clarity and organization of the manuscript, assist with LaTeX formatting, and support debugging. All scientific decisions, experimental design, implementation, interpretation of the results, and final manuscript revisions were performed and verified by the authors, who take full responsibility for the content of this publication.

References

[1] N. Martínez-Guevara, G. Bel-Enguix, V. Soto-Mendoza, J. Beltrán, A. Curiel, S. L. Ojeda-Trueba, M. Ramírez-González, H. Gómez-Adorno, Overview of WomenHelp at IberLEF 2026: Classification of gender-based violence reports to help threatened women in Northern Mexico, Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview ofIberLEF 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] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel 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 (SemEval-2019), 2019, pp. 54–63. doi:10.18653/v1/S19-2007.

[4] E. Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, Transactions of the Association for Computational Linguistics 7 (2019) 677–694. doi:10.1162/tacl_a_00293.

[5] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:10.1613/jair.1.12752.

[6] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2021, pp. 2591–2597. doi:10.18653/v1/2021.naacl-main.204.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.

[8] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC Workshop at the International Conference on Learning Representations (ICLR), 2020.

[10] 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. doi:10.18653/v1/2020.acl-main.747.

[11] 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).

[12] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.