LAB-CO at WomenHelp 2026: Diagnostic-First Ensembling for Spanish Gender-Violence Severity Classification Under Mild Distribution Drift

Angel Serrano[1,*], Thomas Favennec[1], Valentina Carbonell[1] and Lilia Montoya[1]

[1]LAB-CO (Laboratorio de Soluciones Colaborativas), Mexico City, Mexico

Abstract

This paper describes our submission to Subtask 1 of the IberLEF 2026 *WomenHelp-MX* shared task, which consists of classifying Spanish-language institutional reports of gender-based violence from Northern Mexico into four severity levels: Mild, Medium, High, and Severe. Our main contribution is a diagnostic-first model-selection protocol that addresses the distributional shift documented across corpus splits through adversarial validation: rather than relying on development-set performance as a proxy for test performance, we optimize a two-term criterion combining out-of-fold training macro-F1 with a KL-divergence regularizer over predicted test class marginals. Our final system is a five-component soft-voting ensemble trained under this criterion. It obtained first place on Subtask 1 with an official macro-F1 of 0.627.

Keywords

gender-violence detection, Spanish NLP, ensemble learning, adversarial validation, covariate shift, pseudo-labeling, BETO, IberLEF 2026

1. Introduction

Gender-based violence is a public health and justice problem of documented severity in Northern Mexico. Institutional care processes generate written reports that describe the nature, context, and consequences of each incident. When these reports are used to route victims to the appropriate care circuit or to support a legal case, the accuracy of severity classification has direct operational consequences: routing a Severe case as Mild can delay or deny access to specialized protection. Automating this classification at scale requires systems that are not only accurate in aggregate, but robust to the class imbalance and linguistic regularities specific to this corpus.

The *WomenHelp-MX* shared task [1] at IberLEF 2026 [2] operationalizes this problem as a four-class severity classification task (Mild, Medium, High, Severe) over anonymized Spanish-language institutional reports. This paper describes our system for Subtask 1.

Three properties of this corpus drive our modeling decisions:

- **Severe class imbalance with operationally costly minority.** The Severe class represents only 4.5% of training samples, yet corresponds to cases involving sexual violence, attempted homicide, weapons, and abduction, making false negatives on this class operationally far costlier than the inverse error.
- **Surface-level regularities induce tokenization pathologies.** All reports are transcribed in uppercase, which causes a standard cased Spanish BERT tokenizer to over-segment input by nearly a factor of two and truncate 33.7% of the corpus at `max_length=256`; a simple lowercasing step eliminates most of this pathology (Section 4).

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

angel.serrano@lab-co.org (A. Serrano*); thomas.favennec@lab-co.org (T. Favennec); valentina.carbonell@lab-co.org (V. Carbonell); lilia.montoya@lab-co.org (L. Montoya)

0009-0006-3152-6031 (A. Serrano*); 0009-0000-8673-8409 (T. Favennec); 0009-0000-3808-0331 (V. Carbonell); 0009-0005-6618-1873 (L. Montoya)

- **Mild but non-trivial distribution drift across splits.** The three corpus splits exhibit a mild but measurable distributional shift (Section 3.2), which makes development-set F1 a structurally unreliable proxy for test performance.

Our main contribution is a diagnostic-first model-selection protocol designed for shared-task settings under mild covariate shift. Rather than selecting models by development-set F1 alone, we optimize a two-term criterion that combines out-of-fold training macro-F1 with a KL-divergence regularizer over predicted test class marginals (Section 5.1). This protocol directly addresses the reliability problem identified in the corpus analysis. The system built under this protocol, a five-component soft-voting ensemble (Section 5.2–5.3), obtained first place on Subtask 1 with an official macro-F1 of 0.627.

Secondary contributions include: an empirical comparison against four calibration baselines showing none uniformly dominates the proposed criterion on this corpus (Section 6.1); a leave-one-component-out ablation justifying ensemble composition (Section 6.2); a multi-criteria submission gate for iterative-submission risk control (Section 5.5); and a confusion-matrix error analysis with attention to the operational cost of Severe misclassification (Section 8).

2. Background and Related Work

2.1. Spanish Gender-Violence Classification

Detection of gender-based violence and misogyny in Spanish has been driven largely by a series of shared tasks at IberEval and IberLEF. The AMI task at IberEval 2018 [18] introduced binary misogyny detection together with a behavioural taxonomy — stereotype, dominance, sexual harassment, discredit — over English and Spanish tweets, establishing the genre of fine-grained abuse categorisation in Spanish. The EXIST shared-task series [19] extended this lineage to a five-category sexism taxonomy that includes an explicit sexual-violence category, while HatEval (SemEval-2019 Task 5) [20] provided a parallel multilingual benchmark for hate speech against women. For graded severity specifically, the DETOXIS task at IberLEF 2021 [21] is the closest precedent in Spanish: its second subtask classifies comment toxicity on a four-level ordinal scale, a structure directly analogous to our Mild/Medium/High/Severe scheme. Institutional-care narratives such as those in WomenHelp-MX differ from all of the above on at least three axes (register, length, and anonymization-induced lexical regularization) and admit a richer severity ordinal that does not collapse to the binary misogynistic/non-misogynistic schema of much prior work [8]. For Spanish-language transformer encoders, BETO [3] remains a strong baseline; RoBERTuito [4] is its social-media analogue.

Taken together, these shared tasks have established detection and binary or categorical classification as the dominant framing for Spanish GBV text. WomenHelp-MX extends this line of work to institutional narratives and to an explicit four-level severity ordinal, a combination that prior shared tasks have not addressed.

2.2. Imbalanced Ordinal Classification

Focal loss [5] with class-balanced reweighting [6] is a standard recipe for severe-tail classes. Rank-consistent ordinal output layers such as CORAL [22] and CORN [16] are reported to improve on standard cross-entropy when classes admit a strong ordinal structure by enforcing monotonicity constraints across cumulative binary heads. The CLPsych 2019 shared task [23] framed suicide-risk prediction as a four-level ordinal problem (no risk / low / moderate / severe) over social-media posts and demonstrated that class-imbalanced ordinal schemes with operationally asymmetric misclassification costs require targeted loss design. We evaluate CORN as an ablation in Section 5.5.

These methods assume that the ordinal structure of the target variable is the primary modeling challenge. Our setting adds a second constraint: the minority class at the top of the ordinal (Severe) carries an asymmetric operational cost that standard macro-F1 does not capture, motivating the class-balanced focal loss and the Severe-specific bounds in our submission gate.

2.3. Covariate Shift and Adversarial Validation

The general framework for handling distribution drift in supervised learning is given by Sugiyama and Kawanabe [10]; adversarial validation [11] is a tractable empirical diagnostic. We use the AUC of a binary discriminator trained to separate two splits as our drift estimator and treat values close to 0.5 as evidence of distributional similarity. The degradation of development-set reliability under distribution shift is a known failure mode in NLP: Risch and Krestel [24] show that BERT ensembles improve robustness to cross-domain aggression detection, consistent with our finding that soft-voting outperforms learned stackers under train↔test drift; Wiles et al. [25] provide a fine-grained analysis of how model behaviour changes across shift types. For label-shift correction, the standard reference is Lipton et al. [9]; in this paper we do not perform explicit label-shift correction but rather use the KL distance to training priors as a regularizer on submission selection.

Prior work has largely treated distribution shift as a problem to be corrected through reweighting or domain adaptation. We take a complementary approach: rather than correcting for shift, we use adversarial validation as a diagnostic to justify replacing development-F1 with a more conservative model-selection criterion.

2.4. Pseudo-Labeling

Lee's pseudo-labeling [12] remains a useful semi-supervised baseline. More recent work has refined the approach for transformer-based text classification: Unsupervised Data Augmentation (UDA) [26] combines pseudo-labeling with consistency regularization and achieves strong gains in text classification; Uncertainty-aware Self-Training (UST) [27] introduces confidence-based sample selection particularly suited to low-resource and class-imbalanced settings; and Du et al. [28] demonstrate that self-training improves pre-trained language models across a range of NLU benchmarks. We adopt a soft, weighted variant with a confidence threshold and a sub-unitary loss weight, drawing on the selective-threshold intuition of UST, and we never validate against pseudo-labeled samples.

These methods typically assume a stationary test distribution. We apply pseudo-labeling under mild covariate shift, using a confidence threshold to limit the propagation of errors from distribution-shifted test examples into the training signal.

Across all four strands, the common gap is the combination of a formal, high-stakes institutional register, an ordinal severity scheme with asymmetric misclassification costs, and a measurable train–test distribution mismatch that makes development-set evaluation unreliable. The present work addresses all three simultaneously through diagnostic-first model selection, a calibration-anchored ensemble, and a multi-criteria submission gate.

3. Dataset

3.1. Corpus Statistics

The Subtask 1 corpus contains 21,254 anonymized reports distributed as train (13,630, 64.1%), development (3,405, 16.0%), and test (4,219, 19.8%). Class proportions on the training split are reported in Table 1.

The Severe class is critically underrepresented at 4.5% of training samples, yet corresponds to cases involving sexual violence, attempted homicide, weapons, and abduction. This asymmetry motivates the class-balancing strategy described in Section 5.2.

Class	n	%	Indicative content (from data card)
Mild	2,956	21.7	Legal counseling, custody, alimony
Medium	4,738	34.8	Psychological violence, verbal threats, economic control
High	5,320	39.0	Physical violence, death threats, compound violence
Severe	616	4.5	Sexual violence, attempted homicide, weapons, abduction

Table 1: Subtask 1 training-set class distribution. The Severe class is severely under-represented despite carrying the highest operational cost of misclassification.

3.2. Distribution Drift Across Splits

We performed an adversarial-validation diagnostic to assess whether the three corpus splits are drawn from the same distribution. For each ordered pair of splits, we trained a binary discriminator on softmax-normalized BETO OOF logits and report 5-fold cross-validated ROC-AUC under two discriminators: L2-regularized Logistic Regression and LightGBM [15]. Results are in Table 2.

All AUC values are below 0.7, indicating a regime of mild covariate shift. In either discriminator reading, the development split is not an i.i.d. sample from the test distribution, which makes development-F1 a structurally weak proxy for test performance. This finding directly motivates the model-selection criterion proposed in Section 5.1.

Pair	AUC (Logistic)	AUC (LightGBM)
train $\leftrightarrow$ dev	0.515 ± 0.017	0.676 ± 0.011
train $\leftrightarrow$ test	0.584 ± 0.007	0.648 ± 0.007
dev $\leftrightarrow$ test	0.584 ± 0.010	0.632 ± 0.007

Table 2: Adversarial-validation AUC across the three splits using softmax-normalized BETO OOF logits as features. Values close to 0.5 indicate distributional similarity. Both linear and non-linear discriminators consistently identify mild but non-trivial drift; the picture from the two discriminators differs in detail, reflecting that the drift signal is partly non-linear, but agrees that drift is present.

4. Preprocessing

A single deterministic pre-processing decision contributed more to baseline performance than any architectural change reported in this paper, and we believe it is the most directly transferable practical finding of the work. Because the corpus is transcribed entirely in UPPERCASE, BETO's cased wordpiece tokenizer over-segments aggressively: virtually every content word is fragmented at unusual boundaries, producing a wordpiece sequence roughly $1.7\times$ longer than necessary. Figure 1 shows the token-length distribution before and after a single `text.lower()` pass.

Empirically, this change reduced the mean wordpiece count per report from 233 to 133, reduced truncation at `max_length=256` from 33.7% to 8.5%, and improved BETO development macro-F1 from 0.502 to 0.537 (+3.5 points). At the `max_length=512` setting used in this paper, the corresponding truncation rates are 5.4% (UPPERCASE) and 0.04% (lower-cased), effectively eliminating truncation as a confound. All transformer components in this paper consume the lower-cased text. We highlight this finding because it is, in our view, the kind of dataset-specific pre-processing pathology that is invisible to architecture-centric ablation tables but dominates them in practical effect.

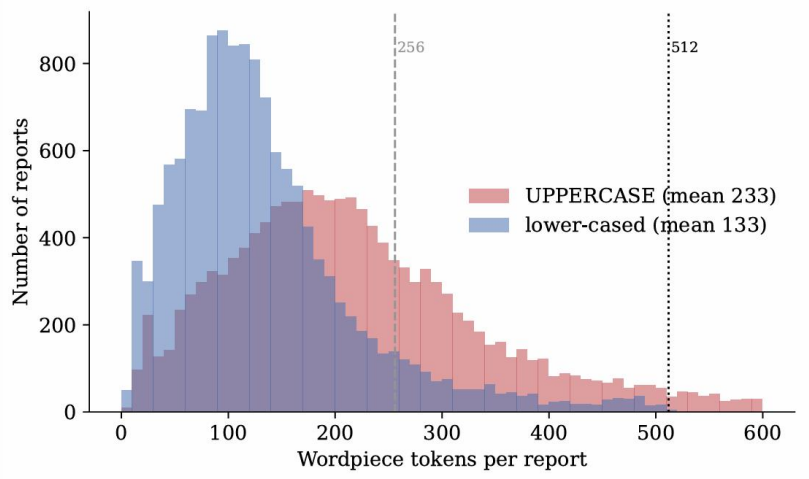

Figure 1: Token-length distribution of the training corpus under BETO's wordpiece tokenizer, before and after lower-casing. UPPERCASE input has mean 233 wordpieces with 33.7% of reports truncated at `max_length=256`; lower-cased input has mean 133 wordpieces with 8.5% truncation at the same limit (and 0.04% truncation at `max_length=512`, the setting used in this paper). Lower-casing alone improved BETO development macro-F1 from 0.502 to 0.537 (+3.5 points), exceeding any subsequent architectural gain in our pipeline.

5. Methodology

5.1. Diagnostic-First Model Selection Criterion

We now formalize the model-selection criterion implied by the diagnostic of Section 3.2. Our framing is deliberately conservative: we do not claim a theoretical proof that development-F1 is dominated by the criterion we propose, but rather argue that the criterion is a more conservative surrogate under the empirically observed drift pattern.

Development-F1 is an unbiased estimator of test-F1 only when $P_{\text{dv}} = P_{\text{te}}$. The adversarial-validation result of Section 3.2 shows that neither equality holds exactly. A second, more concrete failure mode is specific to this work: one of our strongest components, the train+dev-fitted *betotd* model of Section 5.2, is by construction unable to contribute development-F1 estimates. Models that include *betotd* as a substantial weight in an ensemble therefore have leaked predictions on the development split, and their nominal development-F1 numbers overstate generalization. The two-term criterion below is one such signal that does not pass through the development split at all.

We adopt the following two-term submission-ranking criterion. Given a candidate model $\hat{f}$, define

$$\Phi(\hat{f}) \;=\; \alpha \cdot F_1^{\text{OOF}}(\hat{f}; S_{\text{tr}}) \;-\; \beta \cdot D_{\text{KL}}\Big(\hat{\pi}_{\text{te}}(\hat{f}) \,\|\, \pi_{\text{tr}}\Big)\,, \tag{1}$$

where F_1^{OOF} is the 5-fold OOF macro-F1 on the training set, D_{KL} is the Kullback–Leibler divergence, and $\alpha, \beta > 0$ are positive weights. We rank candidate submissions by Φ.

The first term, F_1^{OOF}, is an unbiased estimator of in-distribution generalization performance and is a strong proxy when $P_{\text{tr}} \approx P_{\text{te}}$. The second term is a distributional risk regularizer: it penalizes candidates whose predicted test class marginals diverge from the training-set priors, guarding against pathological prediction-marginal collapse toward the majority class. In practice we applied Φ lexicographically: first filtering candidates by an OOF-F1 threshold, then ranking by KL.

The criterion is a reasonable surrogate only under three regularity conditions: adversarial-validation AUC between train and test must be mild (below approximately 0.65); no concurrent label shift is suspected; and the development split either fails an exchangeability check against test or is structurally

unsuitable for evaluation of part of the ensemble, as is the case here with *betotd*. When these conditions fail, development-F1 with appropriate importance-weighting or explicit label-shift correction is to be preferred.

5.2. System Components

All transformer components share: AdamW optimizer with Layer-wise Learning Rate Decay (factor 0.9) [7], cosine learning-rate schedule with 10% linear warmup, focal loss [5] with $\gamma = 2.0$ and class-balanced α weights [6], bfloat16 mixed-precision autocast, gradient clipping at norm 1.0, and 5-fold stratified cross-validation with seed 42. The class-balanced focal weights are computed as

$$\alpha_c \;=\; \frac{\sum_{c'} n_{c'}}{4\, n_c + \varepsilon}, \qquad \alpha_c \;\leftarrow\; \alpha_c / \min_{c'} \alpha_{c'},$$

yielding $\alpha \approx (1.80, 1.12, 1.00, 8.64)$ for Mild, Medium, High, Severe.

The five components are described below.

Component	Architecture / data	OOF train F1
b8e	BETO [3], 8 epochs, max_len 512, 5-fold OOF	0.7189
cls	Word + char TF–IDF, LogReg $\oplus$ calibrated LinearSVC, 5-fold OOF	0.6814
rt	RoBERTuito [4], max_len 128, 5-fold OOF	0.6583
pseudo	BETO + 47% confident test pseudo-labels (weight 0.5), 5-fold OOF	0.7219
betotd	BETO trained on train$\cup$dev, 5 seeds, 4 epochs, no validation	— (no OOF)

Table 3: Component models and their OOF training macro-F1. The *betotd* component has no OOF estimate because it consumes the development split as training data; its development F1 is deliberately not used for model selection.

BETO 8e (b8e). `dccuchile/bert-base-spanish-wwm-cased` (110M parameters), `max_length=512`, batch size 16 with gradient accumulation 2 (effective batch 32), learning rate 2×10^{-5}, 8 epochs with early stopping at patience 3. Doubling `max_length` from 256 to 512 yielded a +4.4-point OOF improvement, consistent with the post-lower-casing length distribution.

Classical (cls). Word n-grams (1–2, top 80K features by frequency) concatenated with character n-grams (`char_wb`, 2–5, top 200K features) and `sublinear_tf=True`. Per fold, two classifiers are trained: `LogisticRegression` ($C = 1.0$, class-balanced weights) and a sigmoid-calibrated `LinearSVC` (3-fold internal calibration). Per-fold probabilities are averaged in probability space.

RoBERTuito (rt). `pysentimiento/robertuito-base-uncased`, `max_length=128`, batch size 32, 6 epochs. RoBERTuito underperformed BETO on this corpus (OOF F1 0.6583 vs. 0.7189), likely because the violence reports (mean approximately 200 tokens, formal register) differ substantially from the Twitter pre-training distribution. It is retained in the ensemble for diversity (see Section 5.3).

Pseudo-labeled BETO (pseudo). A two-stage semi-supervised procedure. Stage 1: train a base *b8e* $\oplus$ *betotd* 50/50 blend and use it to predict the test set. Retain the top-47% confident samples (max softmax probability ≥ 0.7), yielding 1,976 pseudo-labeled test examples whose class distribution (Mild 454, Med 660, High 775, Sev 87) closely matches the training priors. Stage 2: for each of the 5 folds, concatenate the pseudo-labels to the fold's training data with focal-loss sample weight 0.5; the validation fold consumes only real labels.

BETO Train+Dev (betotd). BETO fine-tuned on the concatenated train∪dev split (17,035 samples), with 5 random seeds in $\{42, 123, 456, 789, 2024\}$, 4 epochs each, and no held-out validation. Test predictions are averaged across seeds in probability space. The model has no OOF estimate by construction, and its development predictions are leaked, so its development F1 must not be used as a generalization signal in this paper or in any downstream weight-selection procedure. The justification for including *betotd* in the ensemble rests on two independent facts: (a) the additional supervision from training on train∪dev rather than train alone is genuine and is the standard rationale for using held-out splits at training time once final hyper-parameters are fixed; (b) its predicted test marginals are the closest match to training priors of any single model in our pipeline (Table 5), which is the most defensible single-number proxy for test calibration we have under the regularity conditions of Section 5.1.

5.3. Architectural Diversity

Table 4 reports the pairwise Cohen's κ [14] between component test predictions. The classical component is the most divergent from BETO ($\kappa = 0.534$), confirming that non-neural features contribute genuine architectural diversity rather than redundant signal.

	b8e	*cls*	*rt*	*pseudo*	*betotd*
b8e	1.000	0.534	0.582	0.884	0.742
cls		1.000	0.517	0.551	0.515
rt			1.000	0.586	0.570
pseudo				1.000	0.753
betotd					1.000

Table 4: Pairwise Cohen's κ between component test predictions. Lower values indicate greater architectural diversity. The classical component is uniformly the most diverse: it has the lowest κ against every other component (range 0.515–0.551), confirming that the non-neural feature space captures error patterns not redundant with BERT-family models. The high κ between *b8e* and *pseudo* (0.884) is expected since the latter is a BETO model trained with augmented data.

5.4. Final Ensemble

Let $P_m \in \mathbb{R}^{|S_{\text{te}}| \times 4}$ denote the row-stochastic softmax output of component m on the test set. The final ensemble is

$$P_{\text{final}} \;=\; \frac{1}{2}\underbrace{\left[\frac{1}{4}\big(P_{b8e} + P_{cls} + P_{rt} + P_{pseudo}\big)\right]}_{\text{OOF blend}} + \frac{1}{2}\,P_{betotd}. \tag{2}$$

Predictions are the row-wise $\arg\max$ of P_{final}. The four OOF components contribute 12.5% each; *betotd* contributes the remaining 50%.

The 50/50 split between the OOF blend and *betotd* reflects two complementary roles. The OOF blend provides leak-free, cross-validated probability estimates from four architectures whose pairwise diversity is documented in Section 5.3. The *betotd* component provides a calibration anchor: as Table 5 shows, its predicted marginals are the closest single-model match to training priors, which is exactly the quantity that our submission criterion in Eq. (1) is designed to track.

We did not perform a grid search over weights against the test set, which would constitute test-set fitting. We did verify that the 50/50 split outperformed 70/30 and 30/70 alternatives *on the OOF training F1*.

A clarifying point on the role of KL in this construction: the proposed ensemble is *not* the configuration that minimizes KL to training priors. The pure-*betotd* model is closer in KL (0.0092) than the v_final ensemble (0.0109), as Table 5 shows directly. The reason we do not submit *betotd* alone is that it is single-model, has no OOF estimate, and trades off the architectural diversity documented in Section 5.3.

The v_final ensemble instead occupies a Pareto-frontier position: it has higher OOF training F1 (0.7412) than any single component, comparable KL to the best single model (0.0109 vs 0.0092), and access to the diversity contribution of the four OOF components. The KL term in Eq. (1) is a regularizer that prevents distribution-collapse failure modes, not a target to be minimized in isolation; the present ensemble is consistent with this view.

	Mild	Medium	High	Severe	KL to π_{tr}
Training priors $\times\lvert S_{\text{te}}\rvert$	915	1,466	1,646	190	—
betotd (5-seed avg.)	923	1,689	1,391	216	0.0092
b8e	861	1,765	1,429	164	0.0112
pseudo	870	1,778	1,412	159	0.0124
rt	1,196	1,555	1,281	187	0.0204
cls	795	1,887	1,462	75	0.0285
v_final (ensemble)	872	1,756	1,409	182	**0.0109**

Table 5: Predicted test class counts for each component model and the final ensemble, against the test-size-scaled training priors. The right-hand column reports the Kullback–Leibler divergence between each predicted marginal and the training priors. *betotd* produces the closest marginals among single models; *cls* severely under-predicts the Severe class. The ensemble does not have the lowest KL on the table (that distinction belongs to *betotd* alone), but combines low KL with the OOF diversity discussed in Section 5.3.

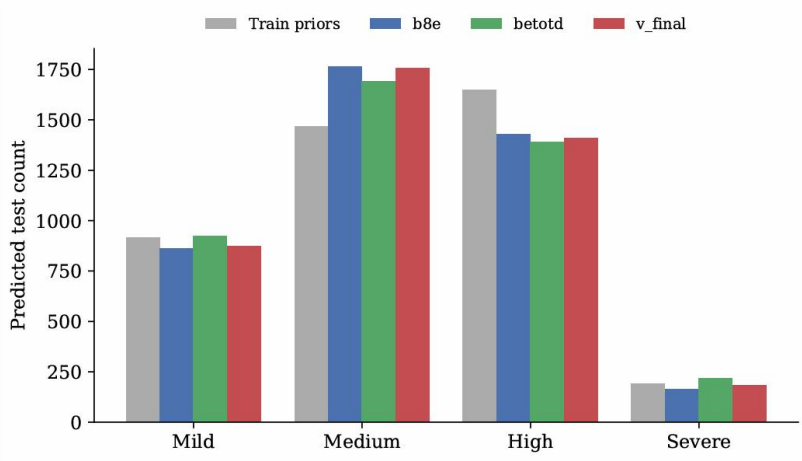

Figure 2: Predicted test class marginals for the strongest single models (*b8e*, *betotd*) and the final ensemble (v_final), against the test-size-scaled training priors. The *betotd* bars are uniformly closest to the priors among single models; the ensemble inherits this calibration anchor while gaining discriminative signal from the OOF components.

5.5. Submission Risk Control

Codabench-style shared tasks impose a hard limit on test-set submissions, creating an iterative risk-control problem. We formalized our policy as a multi-criteria gate: a candidate submission is accepted only if it satisfies all of the following — its OOF training F1 is at least as high as the current best submission; the 10th bootstrap percentile of OOF predictions exceeds 0.71; the KL divergence between predicted test marginals and training priors stays below 0.012; the predicted Severe count falls in $[170, 200]$; and Cohen's κ against the current best lies in $[0.92, 0.97]$. The first two criteria control generalization and variance; the third and fourth enforce distributional health on the test set; the fifth bounds how much the candidate may differ from a proven baseline.

The v_final ensemble satisfies all five criteria with margin: OOF F1 $= 0.7412$, BootP10 $= 0.7349$, KL $= 0.0109$, predicted Severe count $= 182$, and $\kappa = 0.921$ against the proven `v24_07` baseline. In practice, the gate rejected three otherwise plausible submissions that had high development F1 but elevated KL or low Severe counts — a failure pattern consistent with dev-overfit under the drift documented in Section 3.2.

For completeness we report variants that did not improve upon v_final.

Multilingual mDeBERTa-v3-base. OOF F1 0.5706, development F1 0.4170. Substantially weaker than BETO, likely due to mismatch between its broad multilingual pre-training distribution and the Mexican-Spanish formal-register target.

CORN ordinal regression. Implemented on the BETO backbone with three binary cumulative heads. We observed only +0.003 OOF F1, well below the +2 to +4 points reported in the literature. We attribute this to the severity ordinal being weaker than in classical ordinal benchmarks: surface features already discriminate strongly between Mild and Severe, so focal loss captures most of the signal CORN was designed to recover.

Multi-task severity + violence-type BETO. Shared encoder with two heads and joint loss with masking. The auxiliary signal was too sparse to transfer — only 26% of Subtask 1 samples carry Subtask 2 labels. OOF F1 0.7221.

LightGBM meta-stacker. Trained on 28-dimensional OOF features. Best development F1 0.4973, substantially worse than soft-voting. The most plausible failure mode is overfitting to training-set OOF noise patterns that do not transfer under train↔test drift.

Aggressive dev-tuned biases (v4 lineage). Per-class logit bias optimization on development produced test F1 0.61 and saturated quickly — the canonical failure mode of development-overfit under non-exchangeable splits.

6. Experiments and Analysis

6.1. Comparison Against Formal Calibration Baselines

The submission criterion of Eq. (1) and the calibration-anchor role of *betotd* both target the same underlying problem — aligning predicted class marginals with prior expectations — that classical calibration methods address more directly. We benchmark four formal calibration baselines on *betotd* and report their development F1 and KL divergence to training priors on the test set.

Temperature scaling has no effect on argmax-derived metrics: the optimizer drives T to the search boundary, indicating that class-probability ordering is already correct. Isotonic regression appears to improve dev F1 substantially, but this is a within-split overfit — fitted on half of dev, evaluated on the other half — and pushes test marginals away from training priors (KL rises from 0.0092 to 0.0102), making it unsafe as a submission selector. Both BBSE and uniform prior correction degrade dev F1 substantially and produce mis-calibrated test marginals; BBSE fails because it assumes label shift, but *betotd* was trained on train∪dev with the training class distribution, creating a structural mismatch with BBSE's assumptions.

Note that this comparison evaluates calibration methods on a single component, while v_final combines five; a more principled comparison would re-calibrate each OOF component before ensembling, which we leave for future work. Under this caveat, none of the four baselines is a drop-in replacement for the diagnostic-first criterion on this corpus.

Method	Dev F1	KL	Test marginals (M/Med/H/Sev)
betotd (raw)	0.7089	0.0092	923 / 1689 / 1391 / 216
+ Temperature scaling*	0.7089	0.0092	923 / 1689 / 1391 / 216
+ Isotonic regression**	0.7694	0.0102	756 / 1668 / 1674 / 121
+ BBSE prior correction	0.6177	0.4053	352 / 3246 / 400 / 221
+ Uniform prior correction	0.5501	0.1378	1209 / 1477 / 890 / 643
v_final ensemble	0.6570	0.0109	872 / 1756 / 1409 / 182

Table 6: Comparison of *betotd* under four formal calibration methods against the v_final ensemble. *Optimal temperature $T = 5.0$ at the search boundary, indicating that no temperature improves dev F1 (the argmax is invariant under uniform softening). **Per-class isotonic fitted on half of dev, evaluated on the other half; this number is therefore optimistic and not directly comparable to the others, which are computed on full dev.

6.2. Leave-One-Component-Out Ablation

To isolate the contribution of each component, we report a leave-one-component-out (LOCO) ablation on the development split. We note explicitly that the development split is partially leaked through *betotd*, so absolute LOCO numbers should be read with caution; what they reliably establish is the relative contribution of each component.

Removed component	Dev macro-F1	Δ
None (full ensemble)	0.657	$-$
b8e	0.675	$+0.018$
cls	0.648	-0.009
rt	0.660	$+0.004$
pseudo	0.671	$+0.014$
betotd	0.515	-0.142

Table 7: Leave-one-component-out ablation on the development split. *betotd*'s removal causes a 14-point drop in development F1, an order of magnitude larger than any other component; the classical component is the only OOF model whose removal produces a non-trivial degradation (-0.009). The positive Δ values for the BERT-family OOF components (*b8e*, *rt*, *pseudo*) reflect that on the development split, *betotd*'s leaked predictions already encode most of the signal these components provide, so their additional weight is mostly noise in the dev evaluation; this is a strong indication that dev numbers cannot be used in isolation to set ensemble weights.

7. Results

7.1. Submission Trajectory

Table 8 documents the progression of submissions across the competition window, reporting OOF training F1, development F1 with *betotd*-leaked predictions (included for cross-submission comparability only, not used for model selection), KL divergence to training priors, and official test macro-F1. The KL column decreased monotonically while test F1 increased; we report this as an empirical observation on a single trajectory, not as evidence that KL causally predicts test F1.

Submission	OOF F1	Dev F1†	KL	Test F1
v4 (dev-bias-tuned, no *betotd*)	—	0.580	0.056	0.610
v21 (BETO 512 + v4)	—	0.529	0.018	0.610
v22 (*b8e* + 30% *betotd*)	0.730	0.622	0.014	0.620
v_final (Eq. 2)	**0.7412**	0.657	**0.0109**	**0.627**
v_final2 (+4.5% CORN)	0.7425	0.660	0.0101	0.627‡

Table 8: Submission trajectory. †Development F1 reported with *betotd*-leaked predictions, included only for cross-submission comparability; this column is *not* used for model selection. ‡v_final was the official winning submission (macro-F1 0.627); the v_final2 variant passed the internal submission gate but did not yield a separable improvement ($\kappa = 0.968$ to v_final). The KL column decreased monotonically across our actual sequence of submissions while test F1 was non-decreasing; we report this as an empirical observation on a single competition trajectory, not as evidence that KL causally predicts test F1.

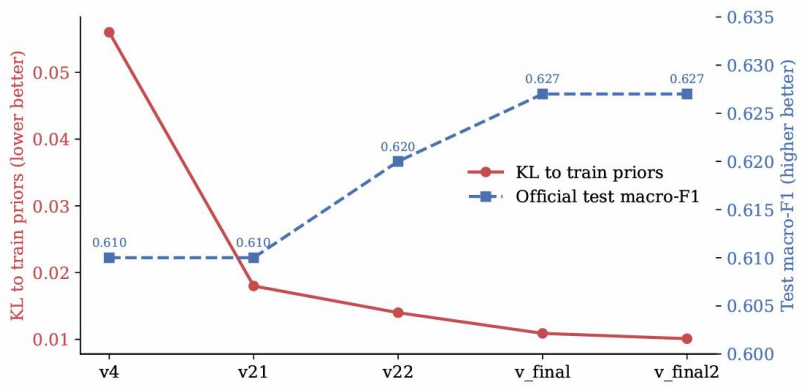

Figure 3: Submission trajectory: KL divergence to training priors (left axis, lower is better) and official test macro-F1 (right axis, higher is better). The two quantities moved in opposite directions across our submission sequence; this is an empirical observation on a single trajectory, not a causal claim about KL predicting test F1.

The system v_final obtained **first place** on Subtask 1 with a macro-F1 of 0.627, with a substantial margin over the runner-up.

7.2. Per-Class Performance

Table 9 reports per-class precision, recall, and F1 on the development split (with *betotd*-leaked predictions, for transparency in disclosing per-class behaviour rather than for model selection).

Class	Precision	Recall	F1	Support
Mild	0.732	0.755	0.744	739
Medium	0.609	0.557	0.582	1,183
High	0.699	0.694	0.696	1,329
Severe	0.496	0.779	0.606	154
macro avg	0.634	0.696	0.657	—

Table 9: Per-class development metrics for the v_final ensemble. The Severe-class recall of 0.779 is achieved at the cost of precision (0.496), reflecting the focal-loss class-balancing and the design choice to over-cover the operationally most consequential class. Reported for transparency; per-class metrics are not used for model selection in this work.

8. Error Analysis

8.1. Confusion Matrix and Asymmetric Operational Cost

Figure 4 reports the row-normalized confusion matrix of the v_final ensemble on the development split. Table 9 shows that the Severe class achieves an F1 of 0.606 in dev, driven by a recall of 0.779 against a precision of 0.496. Aggregated to the confusion-matrix level, the dominant Severe-class failure mode is confusion with the adjacent High class (Severe→High, 18.8% of Severe reports), followed by a much smaller leakage to Medium (1.9%) and a residual 1.3% routed to Mild. The Severe column shows the symmetric cost of high recall: of all predictions labelled Severe (the column), 17.3% are in truth High and 4.1% are in truth Medium.

This is a deliberate operational choice rather than a flaw. In a forensic deployment, a Severe report classified as Mild routes to a legal-counseling stream rather than to an emergency-response stream; a Medium report over-classified as Severe escalates to triage that the report does not need, but does not foreclose appropriate care. The asymmetry between the two error types is large and consequential.

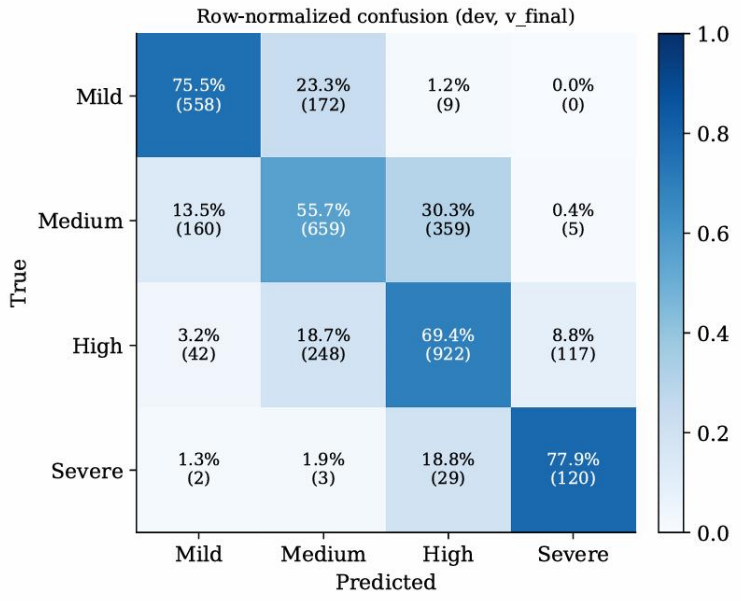

Figure 4: Row-normalized confusion matrix on the development split for the v_final ensemble. Counts are shown in parentheses below the percentage. The Severe row exhibits a 77.9% recall, with the most common confusion being Severe→High (18.8%); only 1.3% of Severe reports are routed to Mild, but this small fraction is operationally the most consequential error mode for a forensic deployment. The Severe column reveals the cost of this recall: 49.6% precision (column-wise Severe predictions include 48.3% of true-High and 2.1% of true-Medium).

Macro-F1 treats these errors symmetrically. A deployment-grade severity classifier should not. Specifically, the cost matrix is not symmetric: false-negatives on Severe are operationally heavier than false-positives on Severe. We do not propose a specific re-weighted metric here, but we flag this as a structural limitation of macro-F1 as the sole optimization target for this shared task.

8.2. Errors by Report Length

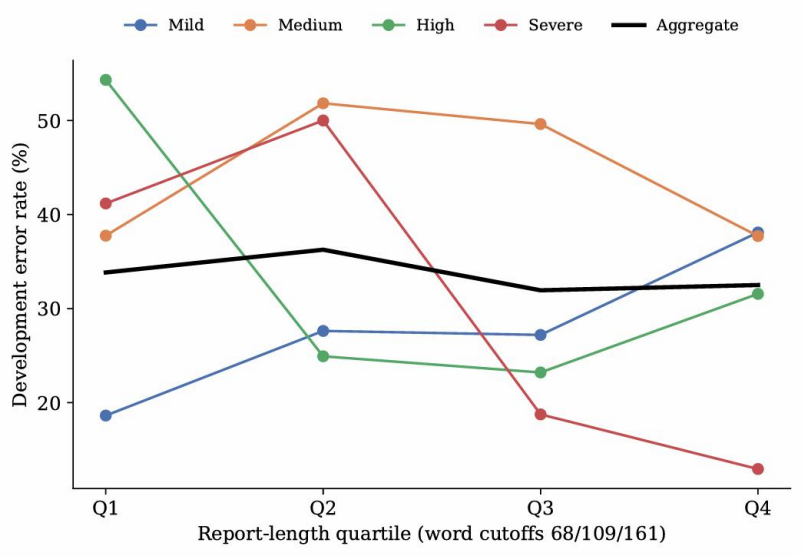

Figure 5: Development error rate by report-length quartile, broken down by true class. The aggregate error rate is essentially flat across quartiles (34.4%, 35.8%, 31.6%, 32.8%), but per-class trends are heterogeneous: the Severe class improves from 41.2% error in Q1 to 12.9% in Q4; the Mild class degrades with length (18.8% to 38.1%).

The aggregate flat trend contradicts the natural intuition that longer reports are harder; the v_final ensemble appears to be largely length-agnostic at the macro level. The per-class breakdown is more diagnostic. The Severe class follows a strongly decreasing error trajectory with length, which we interpret as evidence that long Severe reports contain explicit lexical markers (*abuso sexual*, *tentativa de homicidio*, *arma*) that short reports may omit; this lexical specificity makes long Severe reports easier to classify. Conversely, short High reports are often confused with Severe (column inspection of the confusion matrix on the Q1 subset, not shown), suggesting that brief reports of compound physical violence lack the discriminative context that separates High from Severe. The Mild class shows the opposite, increasing error with length, which we suspect is driven by long Mild reports introducing peripheral mentions of harm that drag predictions toward Medium or High. We did not attempt to verify these hypotheses with token-level attribution, and defer such analysis to future work. We did not perform a second pseudo-labeling iteration; the v_final marginal class distribution on the pseudo-labeled subset (1,976 samples) closely matches the training priors, confirming that the confidence threshold successfully limited distribution-shifted samples from entering the training signal.

9. Discussion

9.1. Why Simple Soft-Voting Outperformed Learned Stackers

We attempted multiple meta-learners (LightGBM, MLP, calibrated Logistic Regression) trained on training-set OOF features, in the tradition of stacked generalization [13]. None matched simple soft-voting on the held-out test split. The plausible cause is overfitting to the noise patterns of training-set OOF predictions, which do not transfer under the train vs. test drift documented in Section 3.2. Soft-voting, by construction, applies uniform weighting and is robust to this failure mode.

9.2. The Role of the Train+Dev-Combined Model

Combining train and development data to fit a single ensemble component is unusual in research settings, where development is preserved for held-out evaluation. It is admissible in competition settings where test predictions are the deliverable, provided one is careful never to use the development metrics of such a model for selection decisions. The ablation in Table 7 shows that *betotd*'s contribution to the development F1 (-0.142 when removed) is an order of magnitude larger than any other component, but this number must be read carefully: the development split is in *betotd*'s training set, so part of this drop is a leakage effect rather than a generalization signal. The complementary evidence is the distribution-level evidence in Section 5.4, which shows that *betotd*'s test marginals are the closest match to training priors among any single model, with KL $= 0.0092$. Under the regularity conditions of Section 5.1, a low KL to training priors is the most defensible single-number proxy for test calibration we have. The combination of these two facts — dev dominance via leakage, and prior-matched test marginals — is what motivates the 50% ensemble weight on *betotd*.

9.3. Generalization of the Diagnostic-First Protocol

We do not claim that the specific weights of Eq. (2) or the specific thresholds of the submission gate transfer to other tasks, and we recognize that the empirical evidence presented here is restricted to a single shared task on a single corpus. With that caveat, we conjecture that the workflow may generalize: for classification tasks where the adversarial-validation AUC between train and test is mild (below approximately 0.65) and where the development split exhibits a non-negligible drift signal against test, the conventional development-F1 optimization paradigm could be productively supplemented by the two-term criterion of Eq. (1). Future work should evaluate this protocol on independent corpora, including other Spanish-language severity tasks, legal or toxicity-ordinal benchmarks, and any shared task whose splits are not constructed by uniform random sampling.

10. Limitations

We identify nine limitations of the present work, each with a natural extension for future work.

- The calibration-baseline comparison of Section 6.1 evaluates four formal methods on the *betotd* component alone; a more principled evaluation would re-calibrate each OOF component before ensembling.
- All claims are supported by a single dataset and competition window, which is too narrow a base for strong generalization claims. Future work should evaluate the diagnostic-first protocol on independent Spanish-language ordinal corpora and on at least one non-Spanish ordinal benchmark.
- Large-language-model zero-shot baselines (Claude, GPT-4, DeepSeek) were not evaluated due to budget constraints. The literature on LLM ensembles for text classification suggests potential gains from a 5–10 LLM ensemble [17].
- All folds use a single random seed (42); multi-seed cross-validation would tighten variance estimates and is a low-cost extension.
- All transformer components shared a single training recipe with minimal per-component hyperparameter search; per-model tuning could yield additional gains.
- As discussed in Section 8.1, macro-F1 is not the right deployment metric for this task. We did not commit to a specific cost-weighted alternative, in part because the shared task itself is evaluated on macro-F1.

- Spanish-specific gender-violence lexicons from prior work were not integrated as features, which might particularly benefit the operationally critical Severe class.
- No data augmentation was explored for the $n = 616$ Severe training samples; back-translation or paraphrase augmentation is a natural follow-up.
- Pseudo-labeling was run for a single iteration; a second iteration with confidence-weighted disagreement filtering is a natural extension.

11. Conclusions

This paper describes our first-place submission to Subtask 1 of *WomenHelp-MX* at IberLEF 2026, with an official test macro-F1 of 0.627. The main contribution is a diagnostic-first model-selection protocol for shared-task settings under mild covariate shift: rather than optimizing development-F1 directly, we propose a two-term criterion that combines OOF training F1 with a KL-divergence regularizer between predicted test marginals and training priors. This criterion directly addresses the problem documented in this corpus, where the conventional development-F1 model-selection paradigm becomes a weak proxy for unobserved test performance.

The system built under this protocol is a five-component soft-voting ensemble. A leave-one-component-out ablation isolates the train+dev-fitted *betotd* component as the dominant ensemble member; its dominance is explained not only by leakage but by its capacity to anchor predicted test marginals close to training priors, which is exactly the quantity the proposed criterion is designed to track. A comparison against four formal calibration baselines shows that none of them is a drop-in replacement for the diagnostic-first protocol on this corpus.

As a secondary finding with direct transferability, we identified a tokenization pathology specific to this corpus: UPPERCASE input causes 33.7% of reports to be truncated under standard `max_length=256` settings. A single lower-casing pass reduces this rate to 8.5% at +3.5 points of macro-F1, exceeding any architectural change reported in this work. Any UPPERCASE corpus processed with a case-sensitive transformer should apply this step before any architectural decision.

Reproducibility

All code (component training scripts, ensemble construction, distribution diagnostics, and submission-gate logic), per-fold OOF logits, and submission CSV files are available on request. Hardware: training conducted on RunPod RTX 3090 and A40 instances. Total compute: approximately 12 GPU-hours at an aggregate cost of $10 USD. All random seeds are fixed at 42 unless otherwise stated.

Declaration on Generative AI

During the preparation of this work, the authors used Anthropic's Claude (Claude 4 family) and OpenAI's GPT-4 family for: *Grammar and spelling check* and *Paraphrase and reword.* After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

We thank the IberLEF 2026 organizers and the WomenHelp-MX shared-task team for curating the dataset and for the evaluation infrastructure. We also thank RunPod for cloud GPU access.

References

[1] N. Martínez-Guevara, G. Bel-Enguix, V. Soto-Mendoza, J. Beltrán, A. Curiel, S. L. Ojeda-Trueba, M. Ramírez-González, and 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, and L. Chiruzzo. Overview of IberLEF 2026: Natural language processing challenges for Spanish and other Iberian languages. In *Proceedings of IberLEF 2026, co-located with SEPLN 2026*. CEUR-WS.org, 2026.

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

[4] J. M. Pérez, J. Giudici, and F. Luque. pysentimiento: A Python toolkit for sentiment analysis and social NLP tasks. *arXiv preprint arXiv:2106.09462*, 2021.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of ICCV*, pages 2980–2988, 2017.

[6] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in NeurIPS*, 2019.

[7] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339, 2018.

[8] F. M. Plaza-del-Arco, M. D. Molina-González, M. T. Martín-Valdivia, and L. A. Ureña-López. An approach based on transformers and topic modeling for the analysis of women's online violence detection in Spanish. *IEEE Access*, 10:78122–78131, 2022.

[9] Z. C. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of ICML*, pages 3122–3130, 2018.

[10] M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments*. MIT Press, 2012.

[11] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[12] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[13] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[14] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in NeurIPS*, 2017.

[16] X. Shi, W. Cao, and S. Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26:941–955, 2023.

[17] A. Kamen and Y. Kamen. Majority rules: LLM ensemble is a winning approach for content categorization. *arXiv preprint arXiv:2511.15714*, 2025.

[18] E. Fersini, P. Rosso, and M. Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of IberEval*, CEUR-WS Vol. 2150, pages 214–228, 2018.

[19] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67:195–204, 2021.

[20] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of SemEval-2019*, pages 54–63, 2019.

[21] M. Taulé, M. A. Martí, F. Rangel, P. Rosso, C. Bosco, and V. Patti. Overview of DETOXIS at IberLEF-2021: DEtection of TOxicity in comments In Spanish. *Procesamiento del Lenguaje Natural*, 67:209–218, 2021.

[22] W. Cao, V. Mirjalili, and S. Raschka. Rank-consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

[23] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. ACL, 2019.

[24] J. Risch and R. Krestel. Bagging BERT Models for Robust Aggression Identification. In *Proceedings of TRAC-2020*, pages 55–61. LREC, 2020.

[25] O. Wiles, S. Gowal, F. Stimberg, S. Shankar, I. Dorian, K. Dvijotham, and A. Cemgil. A Fine-Grained Analysis on Distribution Shift. In *ICLR*, 2022.

[26] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised Data Augmentation for Consistency Training. In *Advances in NeurIPS*, 2020.

[27] S. Mukherjee and A. Awadallah. Uncertainty-aware Self-training for Few-shot Text Classification. In *Advances in NeurIPS*, 2020.

[28] J. Du, Z. Grave, B. Gunel, V. Chaudhary, O. Celebi, M. Auli, V. Stoyanov, and A. Conneau. Self-training Improves Pre-training for Natural Language Understanding. In *Proceedings of NAACL*, pages 5408–5418, 2021.