Detecting and Characterizing Social Support in Social Media: A Multi-Model Approach

Adriana Alejandra Pérez Echeverría[1], Luis Rodrigo Consuelos Figueroa[1] and Moein Shahiki Tash[1]

[1]*Department of Computer Science,*
Master's Program in Data Science,
Universidad Panamericana, Mexico City, Mexico

Abstract

Social support detection refers to the automatic identification and classification of prosocial language in social media, such as expressions of empathy, advice, encouragement, or assistance. This task is commonly framed as a hierarchical classification problem that captures different dimensions of support, including whether a comment is supportive, the target of the support (individual or group), and the specific community being addressed.

In this work, we model social support detection as a three-level hierarchical classification problem across English and Spanish. The three subtasks require systems to determine whether a comment is supportive, whether the support targets an individual or a group, and which specific community is being addressed among six categories: Nation, Other, LGBTQ, Black Community, Religion, and Women.

We evaluate eight model families under 5-fold stratified cross-validation: Logistic Regression, SVM, XGBoost, and Random Forest with TF-IDF features (Tier 1); MLP, TextCNN, and BiLSTM with additive attention using GloVe and FastText embeddings (Tier 2); and fine-tuned transformer models — `roberta-base` for English and both BETO and XLM-RoBERTa for Spanish (Tier 3). Hyperparameters for Tier 1 models are selected via grid search with 5-fold cross-validation. Class imbalance is addressed with SMOTE for traditional models, random oversampling for neural models, and balanced class weights throughout.

Fine-tuned transformers achieve the best results across all subtasks: on the official test set, RoBERTa reaches macro-F1 of 0.7974, 0.8239, and 0.7748 on English Subtasks 1–3, and XLM-RoBERTa reaches 0.8577, 0.8626, and 0.8875 on Spanish Subtasks 1–3. The Spanish track consistently outperforms English, with the largest gap in Subtask 3 (+0.113 macro-F1), driven primarily by a more balanced label distribution in the Spanish corpus. Among non-transformer models, TextCNN with FastText embeddings is the strongest baseline across both tracks. Detailed per-subtask metrics (accuracy, precision, recall, and weighted-F1) are reported in the Results section.

Keywords

social support detection, text classification, class imbalance, RoBERTa, BETO, Traditional Models

1. Introduction

Social media platforms have become primary environments for human communication at scale, enabling billions of users worldwide to share experiences, seek advice, and express solidarity toward others [1]. Beyond content creation and information exchange, these spaces host a broad range of interpersonal interactions with significant emotional and social weight. Among these, prosocial behaviours — such as encouragement, empathy, and collective solidarity — have attracted growing interest from the computational linguistics community, as they reflect communicative dynamics that extend well beyond the affective polarity captured by traditional sentiment analysis [4].

Natural Language Processing (NLP) provides principled tools to analyse social media text automatically and extract actionable signals at scale [2]. However, the informal nature of user-generated content — characterised by slang, contracted forms, missing punctuation, and heavy emoji use — poses persistent challenges for automatic classification systems [3]. Importantly, most prior NLP work on social media has focused on the detection of negative phenomena such as hate speech, offensive language, and misinformation, leaving positive and prosocial communication comparatively understudied [5].

IberLEF 2026, September 2026, León, Spain
adrianaperezech@gmail.com (A. A. P. Echeverría); rodrigocf1250@gmail.com (L. R. C. Figueroa); mshahiki@up.edu.mx (M. S. Tash)

Social support is one such positive phenomenon. Defined as the expression of encouragement, care, admiration, help, or solidarity toward another person or group [8], social support in online communities has been linked to improved well-being and resilience in the face of adversity [6]. From a computational standpoint, social support detection (SSD) requires not only identifying the affective valence of a message but also understanding its interpersonal intent and the identity of its target — whether a specific individual or a broader community. This hierarchical structure makes SSD a richer and more challenging task than binary sentiment classification, with direct applications in mental health monitoring, content moderation, and the computational study of prosocial behaviour in digital communities [7].

This work frames this problem as a three-level hierarchical classification task over social media comments in English and Spanish, as part of the SSD-2026 shared task [27] at IberLEF 2026 [28]. Systems must first determine whether a comment is supportive or not supportive (Subtask 1); if supportive, identify whether the support targets an individual or a group (Subtask 2); and finally, when the target is a group, classify the specific community being supported among six categories: Nation, Other, LGBTQ, Black Community, Religion, and Women (Subtask 3). All subtasks are evaluated under strong class-imbalance conditions using macro-averaged F1, which assigns equal weight to all classes regardless of frequency.

We evaluate a broad spectrum of model families — sparse traditional models with TF-IDF features, static-embedding neural networks (MLP, TextCNN, BiLSTM with attention), and fine-tuned transformer models (`roberta-base` for English; BETO and XLM-RoBERTa for Spanish) — in order to characterise their performance trade-offs under real-world class-imbalance conditions. Our central research question is:

RQ. *Which model family — traditional machine learning, static-embedding neural networks, or fine-tuned transformers — best addresses hierarchical social support detection in English and Spanish social media under class-imbalance conditions, and what performance trade-offs exist between model complexity and classification quality?*

The main contributions of this work are as follows. We compare eight model families under 5-fold stratified cross-validation across both language tracks and all three subtasks, with Tier 1 hyperparameters selected via grid search, and introduce a tiered SMOTE strategy adapted to the extreme class scarcity of English Subtask 3 (*Religion* and *Women*, only 15 training examples each, a 52:1 imbalance ratio). For Spanish we design a language-specific preprocessing pipeline that maps emojis to semantic tokens rather than discarding them, a relevant choice given that 25.3% of Spanish comments contain at least one emoji (vs. 0.6% in English). Our empirical analysis shows that the English–Spanish performance gap is driven primarily by label distribution rather than by model choice or language; a lexical analysis of the most discriminative unigrams and bigrams per class reveals that supportive comments cluster around topical solidarity phrases and emoji-derived tokens while non-supportive comments rely on generic discourse vocabulary, which helps explain why TF-IDF models remain competitive in Subtask 1; and a sentiment analysis (VADER for English, pysentimiento for Spanish) reveals a substantially stronger Supportive/Non-Supportive contrast in Spanish ($+0.68$ compound points) than in English, motivating the inclusion of sentiment scores as auxiliary features in the Spanish traditional pipeline.

The remainder of this paper is structured as follows. Section 2 formally defines the three subtasks. Section 3 reviews related work. Sections 4 and 5 present dataset statistics and corpus analysis. Section 6 describes our methodology. Section 7 details the experimental setup. Section 8 reports results. Section 9 discusses findings and limitations. Section 10 concludes and outlines future work.

2. Task Description

The dataset used in this work [27, 9, 10, 8] poses the automatic identification of prosocial language in social media as a three-level hierarchical classification problem. Each level is formulated as an

independent subtask, with Subtasks 2 and 3 applied only to the subset of comments identified as supportive at the preceding level.

Subtask 1 — Binary Support Detection. Given a social media comment, a system must classify it as either *Supportive* or *Non-Supportive*. A comment is considered supportive when it expresses encouragement, care, admiration, help, or solidarity directed toward a person or group. Comments that openly endorse violence or harm are assigned the *Non-Supportive* label regardless of any otherwise positive surface form.

Subtask 2 — Binary Target Identification. For each comment identified as supportive in Subtask 1, the system must determine whether the expressed support is directed toward a specific *Individual* or toward a *Group*. This distinction requires understanding the scope and social reference of the message rather than its affective tone alone.

Subtask 3 — Fine-Grained Group Classification. For comments classified as group-targeted in Subtask 2, the system must identify the specific community receiving support among six mutually exclusive categories: *Nation*, *Other*, *LGBTQ*, *Black Community*, *Religion*, and *Women*.

All three subtasks are evaluated using macro-averaged F1, which assigns equal importance to every class irrespective of its frequency, penalising systems that achieve high accuracy by concentrating predictions on majority classes.

3. Related Work

Social support detection in online text. The computational study of social support in online communities has developed as a distinct NLP task over the past decade. [9] introduced SSD as a structured NLP problem over a manually annotated dataset of English YouTube comments, framing it as three sequential subtasks and evaluating traditional ML models enriched with linguistic, psycholinguistic (LIWC), emotion, and sentiment features alongside CNN and BiLSTM neural architectures. Their best results using SVM (linear) with combined TF-IDF and LIWC features reached macro-F1 of 0.78 on Subtask 1 and 0.80 on Subtask 2; the best Subtask 3 result (macro-F1 0.73) was achieved by a soft-voting ensemble with TF-IDF. CNN+GloVe embeddings achieved macro-F1 of 0.82 on Subtask 2 and 0.72 on Subtask 3. This work established the task formulation that subsequent studies [10, 8] directly build upon.

[8] extended the same English dataset by evaluating transformer-based models and zero-shot large language models, including GPT-3, GPT-4, and GPT-4o. Fine-tuned `roberta-base` achieved macro-F1 of 0.80, 0.86, and 0.67 on Subtasks 1–3 respectively, consistently outperforming all LLM baselines under zero-shot conditions (GPT-4o: 0.68, 0.82, 0.63). The authors report improvements of 7% in Subtask 2 and 8% in Subtask 3 over the psycholinguistic-feature baselines of the prior work. This finding motivates our choice of RoBERTa as the primary English model. A complementary LIWC-based analysis of supportive language across Spanish and English online communities further characterises these linguistic and psycholinguistic differences [29].

More recently, [10] introduced the first annotated Spanish-language SSD corpus of 3,189 YouTube comments, providing the first bilingual evaluation of the task. They evaluated traditional ML, CNN and BiLSTM architectures, four transformer models, and GPT-4o. On the unbalanced corpus, GPT-4o achieved the best Subtask 1 performance (macro-F1 0.8531), while the `robertuito-sentiment-analysis` transformer reached macro-F1 of 0.8894 on Subtask 2 and 0.8361 on Subtask 3 when trained on a GPT-4o-balanced dataset. Their finding that class imbalance handling is critical for minority-class performance, and that GPT-4o paraphrase-based augmentation substantially improves Subtasks 2 and 3, directly informs our SMOTE and oversampling design choices.

Hierarchical text classification in social media. Social support detection is structurally related to other hierarchical NLP tasks in which a coarse binary decision gates finer-grained classification. SemEval-2023 Task 10 on Explainable Detection of Online Sexism (EDOS) [11] exhibits the same pattern: macro-F1 drops sharply from the binary level (0.8746) to its 11-class level (0.5606) as label granularity and class imbalance increase, mirroring our own Subtasks 1–3 results. More broadly, the hierarchical structure of SSD introduces error propagation, since misclassifications in Subtask 1 cascade into the evaluation pools for Subtasks 2 and 3 [4].

Class imbalance in text classification. Class imbalance is a persistent challenge in social media NLP datasets. [5] conducted a comprehensive survey of methods for addressing class imbalance in deep-learning-based NLP, finding that no single technique dominates across settings and that oversampling effectiveness depends strongly on dataset characteristics. [6] further benchmarked SMOTE and thirty of its variants on transformer-embedded text datasets, confirming that minority-class underrepresentation consistently impairs learning. These findings motivate our tiered SMOTE strategy for English Subtask 3, where the extreme scarcity of *Religion* and *Women* (15 examples each, 52:1 imbalance) makes standard interpolation-based oversampling unreliable. For neural models, we use random text duplication rather than feature-space interpolation, following the recommendation that SMOTE is poorly suited to high-dimensional embedding spaces [5].

Transformer models for multilingual and Spanish NLP. The dominant paradigm for text classification since 2019 has been fine-tuned transformer models pre-trained on large corpora [15]. For Spanish-language tasks, two pre-trained models are particularly relevant. BETO [16] is a BERT-base model trained exclusively on Spanish text using whole-word masking, offering strong in-domain representations for Spanish social media. XLM-RoBERTa (XLM-R) [17] is a multilingual model trained on 100 languages, providing robust cross-lingual transfer. We evaluate both on the Spanish track to assess whether language-specific pre-training or broad multilingual exposure yields better representations for this task.

Emoji handling in Spanish social media NLP. A distinctive feature of the Spanish corpus is its high emoji density: 25.3% of Spanish comments contain at least one emoji, compared to just 0.6% in English. Standard preprocessing pipelines discard emojis as non-ASCII noise, a strategy appropriate for low-emoji corpora but harmful when emojis carry class-discriminative information. We address this by mapping emojis to semantic Spanish-language tokens before applying the base pipeline, preserving their signal for all model families. To our knowledge, semantic emoji mapping as a preprocessing decision in the context of SSD has not been previously examined [9, 10, 8].

4. Dataset Statistics

The shared task provides two parallel annotated corpora, one in English and one in Spanish, each labelled with the three-level hierarchy described in Section 2.

4.1. English Dataset

The English training set originally contains 7,998 social media comments. After quality cleaning — removing 104 duplicate texts with consistent labels and discarding 19 texts with conflicting annotations — the working dataset contains **7,856 comments**. Table 1 summarizes the class distribution across all three subtask levels.

Table 1: Class distribution in the English training set (after deduplication). *Ratio*: majority:minority per subtask.

Subtask	Class	N	%	Ratio
Subtask 1	Non-Supportive	6,084	77.4	3.4:1
	Supportive	1,772	22.6	
Subtask 2	Group	1,444	81.5	4.4:1
	Individual	328	18.5	
Subtask 3	Nation	784	54.3	1.0:1
	Other	416	28.8	1.9:1
	LGBTQ	123	8.5	6.4:1
	Black Community	91	6.3	8.6:1
	Religion	15	1.0	52.3:1
	Women	15	1.0	52.3:1

Comment length follows a right-skewed distribution typical of social media: mean 18.6 words (std 16.3), median 13, range 1–157 words. The corpus is thematically concentrated around the Israeli–Palestinian conflict, the Black Lives Matter movement, and LGBTQ+ rights, which largely determines the class boundaries in Subtasks 2 and 3 and partly explains why simple lexical models remain competitive.

4.2. Spanish Dataset

The Spanish training set contains **2,600 comments**. The Spanish corpus presents a markedly different Subtask 3 structure compared to English: while the English data contains only 15 training examples each for *Religion* and *Women* (52:1 imbalance), the Spanish corpus has a maximum ratio of only 9.3:1 (Black Community vs. LGBTQ). Furthermore, the dominant Subtask 3 class in Spanish is *LGBTQ* (46.2% of group comments), whereas in English it is *Nation* (54.3%), reflecting the different thematic composition of the two corpora. Table 2 summarizes the Spanish class distribution.

Table 2: Class distribution in the Spanish training set. *Ratio*: majority:minority per subtask.

Subtask	Class	N	%	Ratio
Subtask 1	Non-Supportive	2,010	77.3	3.4:1
	Supportive	590	22.7	
Subtask 2	Group	456	77.3	3.4:1
	Individual	134	22.7	
Subtask 3	LGBTQ	224	46.2	1.0:1
	Other	82	16.9	2.7:1
	Religion	60	12.4	3.7:1
	Women	48	9.9	4.7:1
	Nation	47	9.7	4.8:1
	Black Community	24	4.9	9.3:1

This structural difference in label distribution is the primary factor behind the substantially higher macro-F1 scores on Spanish Subtask 3, as discussed in Section 9.

5. Corpus Analysis

5.1. Lexical Analysis

Examining the most frequent unigrams, bigrams, and trigrams per class shows that lexical models capture a mix of topical, affective, and stance signals. In English, supportive comments cluster around solidarity and empathy markers (*free palestine*, *red heart*, *cry face*, *faceloudly cry*, the last being a

transliteration of the crying-face emoji), confirming that emoji-derived signals survive even in the English corpus; non-supportive comments instead carry more descriptive or argumentative vocabulary, with overlapping topical references (*life matter*, *black people*, *black life*). Because this separation is partly topical rather than purely affective, TF-IDF models remain competitive in Subtask 1 (Figures 1 and 2).

In Spanish the separation is less clean: supportive comments use identity, pride, and celebration expressions (*ser bisexual*, *feliz día*, *orgulloso ser*), while non-supportive comments reuse identity vocabulary in definitional or argumentative constructions (*hombre mujer*, *ser hetero*, *orientación sexual*) — so the classes are distinguished by how concepts are framed rather than by topic alone. The same pattern holds for Subtasks 2–3, where coarser labels associate with broad support and stance markers and finer-grained labels depend on narrower topic-specific cues (Palestine terms for Nation, Black Lives Matter terms for Black Community, LGBTQ identity terms, religious vocabulary for Religion).

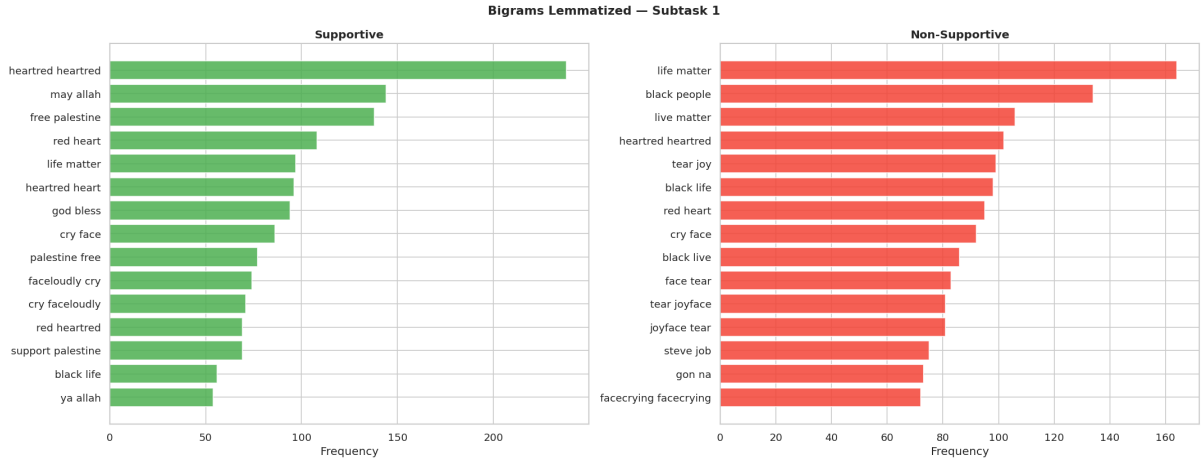

Figure 1: Top bigrams per class, Subtask 1 — English. Supportive comments are characterised by solidarity phrases, heart tokens, and emoji-derived emotion markers (*faceloudly cry*, *red heart*); non-supportive comments include overlapping topical references (*life matter*, *black people*) and generic discourse markers.

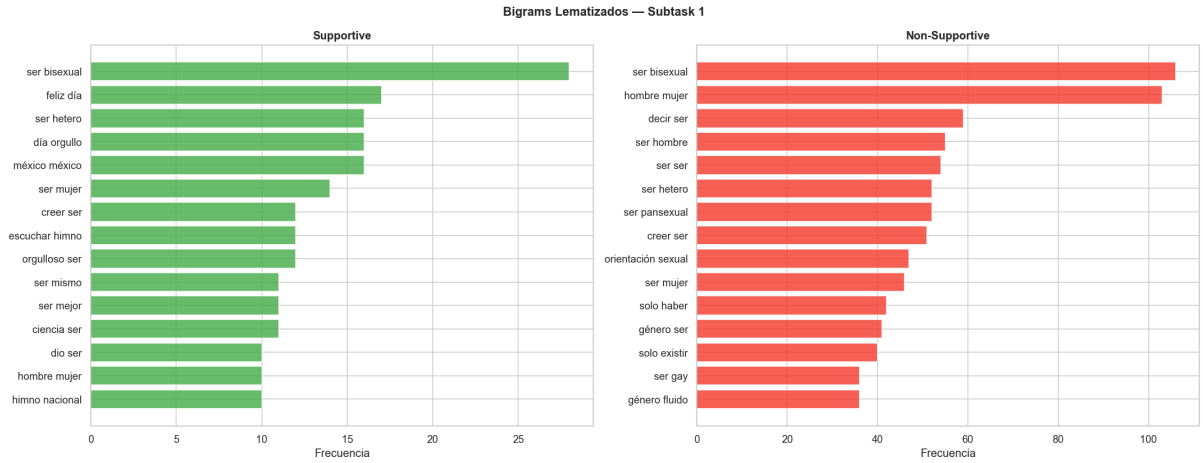

Figure 2: Top bigrams per class, Subtask 1 — Spanish. Supportive comments cluster around identity, pride, and celebration vocabulary (*ser bisexual*, *feliz día*, *orgulloso ser*); non-supportive comments use identity-related vocabulary in definitional or argumentative constructions (*hombre mujer*, *ser hetero*, *orientación sexual*).

5.2. Sentiment Analysis

Sentiment was analysed per corpus with language-appropriate tools: VADER [23] for English and pysentimiento [24] for Spanish. In English, VADER shows only a modest, overlapping difference

between Subtask 1 classes — both produce a KDE peak near $+0.9$, since many non-supportive comments are topically neutral rather than negative — with LGBTQ-targeted comments scoring highest among Subtask 3 categories ($+0.643$; Figures 3 and 5). In Spanish the contrast is far sharper (Figures 4 and 6): the mean compound score is $+0.405$ for supportive versus -0.277 for non-supportive comments (a 0.68-point gap), and Individual-targeted support scores highest overall ($+0.732$). This stronger, cleaner separation — reflecting both the expressiveness of Spanish supportive language and pysentimiento's deeper pragmatic modelling — is why we append pysentimiento scores as auxiliary features to the Tier 1 and Tier 2 Spanish Subtask 1 models, but not to the transformers, which already encode affect through pre-training.

5.3. Text Length Distribution

Token-count distributions before and after preprocessing (Figures 7 and 8) confirm that preprocessing preserves the right-skewed shape. In English the median stays at 13 tokens across the raw, neural, and traditional pipelines (a 1.6% mean reduction); in Spanish, emoji-to-token mapping raises the neural median slightly (13→14) while preserving class-discriminative signals, and the traditional pipeline drops the median to 8 tokens through aggressive stopword removal, reflecting Spanish morphological richness.

6. Methodology

6.1. Preprocessing

English pipeline. The *base pipeline* (applied to all model families) performs: (1) lowercase conversion; (2) informal contraction expansion (*dont* → *do not*, *gonna* → *going to*, etc.); (3) URL removal; (4) user mention and hashtag removal; (5) non-ASCII removal (emojis, Arabic script, Hindi script); (6) punctuation removal; (7) whitespace normalisation. The *traditional pipeline* (TF-IDF models only) additionally performs (8) English stopword removal (198 tokens, NLTK) and (9) WordNet lemmatisation with POS tagging.

Spanish pipeline. An equivalent pipeline is applied to the Spanish data, with two key differences. First, emojis are mapped to semantic Spanish tokens via a custom dictionary (*e.g.*, `bandera_lgbtq`, `corazon`, `apretonde_manos`) *before* the non-ASCII removal step, so that the semantic tokens are retained throughout the pipeline. Second, spaCy's Spanish model [25] is used for tokenisation and lemmatisation instead of NLTK/WordNet, providing better coverage of informal Spanish morphology.

For Subtask 1 specifically, pysentimiento compound sentiment scores are appended as auxiliary input features to both Tier 1 and Tier 2 models: for traditional models, the sentiment vector is concatenated to the TF-IDF representation via sparse matrix stacking; for neural models, it is concatenated to the mean-pooled embedding before the classification head. Transformer models (BETO and XLM-R) are not augmented with external sentiment features, as their contextual pre-training already encodes affective signals.

6.2. Feature Extraction

TF-IDF. Character-level unigrams and bigrams (n-gram range 1–2), `min_df`$= 2$, sublinear TF scaling. The vectoriser is re-fitted within each CV fold to prevent leakage.

Word embeddings for English. Pre-trained GloVe Twitter embeddings (100d and 200d) [12] and FastText Wikipedia English (100d). A vocabulary of up to 30,000 tokens is built from the training corpus; out-of-vocabulary words are initialised with $\mathcal{N}(0,\, 0.1)$. Sequences are padded or truncated to $\text{MAX\_LEN} = 64$ tokens. The embedding layer is kept trainable to allow domain adaptation.

Word embeddings for Spanish. FastText Common Crawl Spanish (CC-ES, 100d), which provides broad coverage of informal web text in Spanish. The same vocabulary and padding strategy as for English are applied.

Transformer tokenisation. `roberta-base` uses byte-pair encoding with a maximum of 128 subword tokens. For Spanish, both `dccuchile/bert-base-spanish-wwm-cased` (BETO) and `xlm-roberta-base` (XLM-R) use their respective sentencepiece tokenisers with the same 128-token limit.

6.3. Models

We evaluate models organised into three tiers by architecture complexity. Tiers 1–2 are identical across both language tracks; Tier 3 differs per language. Hyperparameters for all Tier 1 models were selected via grid search with 5-fold cross-validation on the training set before the main evaluation runs (Section 7).

6.3.1. Tier 1: Sparse Traditional Models

These models operate on TF-IDF representations and require no GPU.

Logistic Regression (LR) is a multinomial logistic regression model with L_2 regularisation, trained using the L-BFGS solver for up to 1,000 iterations with balanced class weights. Grid search identified $C = 10.0$ as the best regularisation strength across all subtasks and both languages.

Support Vector Machine (SVM) is a LinearSVC wrapped in `CalibratedClassifierCV` to produce calibrated probability estimates with balanced class weights. Grid search selected $C = 5.0$ for most subtasks, falling to $C = 1.0$ for Spanish Subtask 2.

XGBoost [18] is a gradient-boosted decision tree ensemble. Grid search optimised the number of estimators (300–500), tree depth (4–8), learning rate (0.05–0.2), and subsample ratio (0.8). Best configurations used 300 estimators with depth 8 and learning rate 0.2 for English Subtasks 1–2, and depth 4 with learning rate 0.05 for Spanish.

Random Forest (RF) is a decision tree ensemble with balanced class weights. Grid search explored tree depth (5 to unlimited) and minimum samples per leaf (1–2), with the best configurations consistently using unlimited depth and 300 estimators.

6.3.2. Tier 2: Neural Models with Pretrained Embeddings

These models are implemented in TensorFlow/Keras and use pretrained word embeddings as input representations.

The **Multilayer Perceptron (MLP)** applies global average pooling over the token embeddings to obtain a fixed-size sentence representation, which is then fed through two fully connected layers (256 and 128 units respectively) with dropout (rate 0.4 after each layer) and a softmax output.

The **TextCNN** [13] applies three parallel one-dimensional convolutional filters with kernel sizes 2, 3, and 4 (128 filters each) to the embedding sequence. Each filter is followed by global max-pooling, and the three resulting feature vectors are concatenated before a dropout layer (rate 0.5) and a softmax output. This multi-scale design captures discriminative n-gram patterns of varying lengths simultaneously.

The **BiLSTM with Additive Attention** [14] processes the embedding sequence with a bidirectional LSTM (128 units per direction) and a spatial dropout layer (rate 0.4) at the input. An additive attention mechanism computes a weighted sum of the LSTM hidden states, producing a context vector that is passed through a final dropout layer (rate 0.4) before the softmax output.

6.3.3. Tier 3: Pre-trained Transformers

For English, we fine-tune `roberta-base` [15] end-to-end. A linear classification head is placed on top of the `[CLS]` token representation, with a dropout layer (rate 0.3) before the output projection.

Training uses AdamW with a learning rate of 2×10^{-5}, weight decay of 10^{-5}, linear warmup over the first 10% of training steps, and gradient clipping at 1.0.

For Spanish, we fine-tune two models with the same classification head and training procedure. **BETO** [16] is a BERT-base model pre-trained exclusively on Spanish text using whole-word masking, offering strong in-domain representations for Spanish social media. **XLM-RoBERTa (XLM-R)** [17] is a multilingual model pre-trained on 100 languages with a RoBERTa-style masked language modelling objective. We evaluate both to assess whether language-specific pre-training or cross-lingual transfer gives better results on this task.

6.4. Class Imbalance Handling

All models apply `class_weight='balanced'` (or its equivalent per framework). Traditional models additionally apply SMOTE [19] within each training fold on the TF-IDF representation, targeting a 2:1 majority:minority ratio for Subtasks 1–2 in both languages.

For English Subtask 3, the extreme per-class imbalance requires a tiered strategy:

- *Nation, Other* (ratio ≤ 2): no SMOTE applied;
- *LGBTQ, Black Community* (ratio ≤ 10): SMOTE to 2:1 vs. Nation;
- *Religion, Women* (ratio ≈ 52): SMOTE to $\min(3 \times n_{\min},\ 0.3 \times n_{\text{Nation}})$.

For Spanish Subtask 3, the maximum imbalance ratio is 9.3:1 (Black Community vs. LGBTQ), which does not require a tiered strategy; standard SMOTE at a 2:1 target ratio is applied uniformly across all minority classes.

The k-neighbours parameter is set to $\min(5,\ n_{\min} - 1)$ to avoid errors with very small classes. Neural models apply random text duplication within each training fold with the same target ratios rather than feature-space interpolation [5].

7. Experimental Setup

Evaluation protocol. All development results are reported under 5-fold stratified cross-validation (80/20 splits), with mean $\pm$ standard deviation macro-F1 over five folds. Oversampling is applied only to the training portion of each fold. The model with the highest development CV macro-F1 per subtask is retrained on the full training set and used to generate predictions for the official test set.

Transformer training. Early stopping with patience of 2 epochs on fold-level macro-F1 is applied during cross-validation, with a maximum of 20 training epochs and a batch size of 32. Final retraining on the full training set uses the same settings with a held-out validation split of 10%.

Software. Preprocessing used NLTK [26] (English stopwords and WordNet lemmatisation), spaCy [25] (Spanish tokenisation and lemmatisation), VADER [23] (English sentiment analysis), and pysentimiento [24] (Spanish sentiment analysis). Traditional models used scikit-learn [20] and XGBoost [18]. SMOTE was implemented via imbalanced-learn [22]. Neural models (MLP, TextCNN, BiLSTM) were implemented in TensorFlow/Keras. Transformer models used PyTorch with the HuggingFace Transformers library [21].

Hardware. All experiments were executed on Google Colab Pro with an NVIDIA T4 GPU (16 GB VRAM). Approximate wall-clock training time per CV fold: 30–60 s (traditional models), 2–5 min (MLP, TextCNN), 5–10 min (BiLSTM), 15–30 min (transformers).

8. Results

8.1. Cross-Validation Results — English

8.1.1. Subtask 1 — Binary Support Detection (EN)

RoBERTa achieves the highest macro-F1 (0.7997), followed by Logistic Regression (0.7878) and TextCNN with GloVe 200d (0.7870). The competitive LR result reflects the topical concentration of the corpus, where discriminative vocabulary (*love*, *palestine*, *children* for Supportive vs. *said*, *know*, *want* for Non-Supportive) maps cleanly to TF-IDF features.

8.1.2. Subtask 2 — Target Identification (EN)

Subtask 2 shows the clearest tier separation: all Tier 2 neural models (0.79–0.81) substantially outperform Tier 1 models (0.68–0.76), and RoBERTa improves further. The Individual/Group distinction requires contextual understanding beyond surface vocabulary, explaining the advantage of sequence-aware architectures.

8.1.3. Subtask 3 — Group Classification (EN)

Subtask 3 is the most challenging for the English track, with high variance across all models (std ≈ 0.04–0.10) due to extreme imbalance in *Religion* and *Women* ($n = 15$ each, ≈ 12 examples per training fold). Table 8 reports overall macro-F1; Table 3 shows per-class F1 for the top models. *Nation* and *LGBTQ* are well-handled across all model families; *Religion* achieves only 0.397 macro-F1 even with RoBERTa, and SVM scores 0 on that class, confirming the near-intractable nature of the extreme imbalance.

Table 3: Per-class macro-F1 (mean, 5 folds), Subtask 3 — English. Best per class in bold.

Model	Nat.	Oth.	LGBTQ	Black	Relig.	Women
RoBERTa	**.907**	**.767**	**.956**	.852	.397	**.752**
TextCNN [FT 100d]	.880	.760	.940	.860	**.420**	.700
Log. Reg.	.869	.720	.948	.832	.347	.666
XGBoost	.893	.762	.906	**.863**	.180	.686
SVM	.868	.707	.943	.811	.000	.560

8.2. Cross-Validation Results — Spanish

8.2.1. Subtask 1 — Binary Support Detection (ES)

Both Spanish transformers substantially outperform all baselines. XLM-R (0.8430) edges BETO (0.8395) by a narrow margin. Logistic Regression (0.8181) remains the strongest non-transformer model, consistent with the English track.

8.2.2. Subtask 2 — Target Identification (ES)

XLM-R (0.8617) achieves the highest macro-F1, with BETO (0.8305) second. TextCNN (0.8026) is the strongest non-transformer baseline, outperforming all Tier 1 models by more than 7 percentage points. XGBoost performs below the other Tier 1 models (0.6259), likely due to the small size of the Individual class combined with the relatively small overall Spanish training set.

8.2.3. Subtask 3 — Group Classification (ES)

Spanish Subtask 3 shows a striking contrast with the English track: XLM-R (0.8585) and BETO (0.8380) far surpass all baselines and achieve substantially higher scores than their English counterparts (0.7718).

Unlike English, *Religion* and *Women* are well-handled by XLM-R (0.891 and 0.901 respectively), reflecting the more balanced Spanish label distribution. Table 4 shows per-class F1 for both transformers and the best traditional baselines.

Table 4: Per-class macro-F1 (mean, 5 folds), Subtask 3 — Spanish. Best per class in bold.

Model	Nat.	Oth.	LGBTQ	Black	Relig.	Women
XLM-R	**.822**	.756	.954	**.953**	**.891**	**.901**
BETO	.754	**.767**	**.960**	.909	.866	.875
Log. Reg.	.694	.630	.908	.671	.815	.717
XGBoost	.724	.618	.858	.747	.754	.615
SVM	.637	.632	.871	.650	.761	.634

8.3. Official Test Results

RoBERTa was selected for all English subtasks and XLM-R for all Spanish subtasks based on CV results. Each model was retrained on the full training set with the same hyperparameters used during development. Table 5 reports official macro-F1 and additional metrics on the held-out test set released by the task organizers.

Table 5: Official test set results. Macro-F1 is the primary evaluation metric. All metrics are computed on the held-out test set released by the task organizers.

Track	Metric	ST1	ST2	ST3
English (RoBERTa)	Macro-F1	**0.7974**	**0.8239**	**0.7748**
	Accuracy	0.8615	0.8980	0.8540
	Macro-Precision	0.7991	0.8255	0.8977
	Macro-Recall	0.7957	0.8223	0.7495
	Weighted-Prec.	0.8607	0.8975	0.8659
	Weighted-Recall	0.8615	0.8980	0.8540
	Weighted-F1	0.8611	0.8977	0.8540
Spanish (XLM-R)	Macro-F1	**0.8577**	**0.8626**	**0.8875**
	Accuracy	0.8986	0.9108	0.9174
	Macro-Precision	0.8545	0.9038	0.8725
	Macro-Recall	0.8611	0.8348	0.9075
	Weighted-Prec.	0.8996	0.9097	0.9239
	Weighted-Recall	0.8986	0.9108	0.9174
	Weighted-F1	0.8991	0.9067	0.9186

The Spanish system outperforms English across all subtasks, with the largest gap in Subtask 3 ($+0.113$ macro-F1 on the official test set, comparing 0.8875 vs. 0.7748). English test scores are very close to their CV estimates (ST1: -0.002, ST2: -0.005, ST3: $+0.003$), indicating strong generalisation with no signs of overfitting. Spanish test scores closely match CV estimates in Subtasks 1 and 2, and the Subtask 3 result (0.8875) falls within the CV standard deviation (0.8585 ± 0.0441), confirming consistent model behaviour across development and test partitions.

9. Discussion

Spanish outperforms English across all subtasks. The most prominent finding is the consistent advantage of the Spanish system (Table 5). For Subtask 3, the gap reaches 0.113 macro-F1 points on the official test set. This difference stems primarily from label distribution: while English Subtask 3

contains only 15 training examples each for *Religion* and *Women* (52:1 imbalance), the Spanish corpus has a maximum imbalance ratio of 9.3:1, enabling XLM-R to achieve strong per-class F1 across all six categories (Table 4). This confirms that the performance gap is a dataset property rather than a model or language property.

XLM-R vs. BETO for Spanish. XLM-R and BETO perform similarly in CV, with differences of 2–3 percentage points. For this corpus — whose dominant topics (Palestine, BLM, LGBTQ+ rights) appear extensively in XLM-R's multilingual training data — broad multilingual pre-training appears equally effective as Spanish-specific pre-training, suggesting that topic-domain match may matter more than language-domain match for this task. Because these differences are smaller than the cross-fold standard deviations in every subtask (e.g., Subtask 3: 0.8585 ± 0.0441 for XLM-R vs. 0.8380 ± 0.0412 for BETO), we do not interpret the XLM-R advantage as statistically reliable; the two transformers are best regarded as comparable on this corpus, and our selection of XLM-R for the official runs reflects its marginally higher mean rather than a significant gap.

Transformer vs. baseline models. Transformers consistently outperform all baselines. The advantage is clearest in Subtask 2, where contextual reasoning beyond surface vocabulary is required. In Subtask 1, Logistic Regression remains within 2–4 percentage points of transformers in both languages, reflecting strong lexical separability driven by the topical concentration of the corpora.

TextCNN as a practical alternative. TextCNN is the strongest non-transformer baseline across most subtasks and both tracks (≈2–5 min training vs. 15–30 min for transformers). Its multi-scale convolutions capture discriminative n-gram patterns effectively. For resource-constrained deployment, TextCNN offers the best efficiency–accuracy trade-off.

Extreme imbalance in English Subtask 3. Even RoBERTa achieves only 0.397 macro-F1 on *Religion* in CV, and SVM scores 0.000. With only 15 training examples, SMOTE at $k = 1$ generates highly extrapolated synthetic examples, and random duplication simply repeats the same 15 data points. Future work could explore LLM-based data augmentation [10] or cross-lingual transfer from the Spanish training data, where the same categories are far better represented.

Topical bias and generalisation. Both corpora are thematically concentrated around a small number of topics. This creates a risk that models learn topic proxies rather than genuine support semantics, which could limit generalisation to out-of-domain social media discussions on different political or social issues.

Further limitations. Three additional caveats apply. First, our system treats the three subtasks independently rather than as a strict cascade; in a real deployment, errors at Subtask 1 would propagate to Subtasks 2 and 3, reducing effective performance beyond what our subtask-level metrics suggest. Second, the sentiment tools are imperfect proxies: VADER was designed for English social media and may not capture the full pragmatic range of supportive language, while pysentimiento, though more powerful for Spanish, is a general-purpose tool not trained on social support data, so sentiment scores should be treated as auxiliary signals rather than definitive affect indicators. Third, all experiments were conducted on a single NVIDIA T4 GPU via Google Colab Pro, and results may vary with different hardware or larger batch sizes, particularly for transformer fine-tuning.

Methodological scope. As a shared-task system-description paper, our primary aim is a controlled empirical comparison of existing model families under a common evaluation protocol rather than a new general-purpose method. The task-specific design choices we introduce — the tiered SMOTE strategy for extreme class imbalance and the semantic emoji-mapping pipeline for Spanish — are nonetheless intended as transferable contributions for related social-media classification tasks.

10. Conclusion and Future Work

We presented a comparison of model families for hierarchical social support detection across two language tracks (English and Spanish) and three subtasks [9, 10, 8]. The main findings are:

1. Fine-tuned transformers are the strongest models: RoBERTa for English (official macro-F1: 0.7974, 0.8239, 0.7748) and XLM-R for Spanish (0.8577, 0.8626, 0.8875). The Spanish system outperforms English in all subtasks, largely due to the more balanced label distribution in Spanish Subtask 3.
2. XLM-R and BETO perform comparably on Spanish (2–3 pp differences). Both far exceed all non-transformer baselines.
3. TextCNN with FastText embeddings is the strongest non-transformer baseline, offering competitive macro-F1 at a fraction of the compute cost and ranking second on most subtasks in both tracks.
4. Logistic Regression with TF-IDF bigrams remains a strong baseline for Subtask 1 in both languages, reflecting topical corpus concentration and the effectiveness of grid-search-tuned regularisation.
5. Extreme class imbalance in English Subtask 3 (52:1 for *Religion* and *Women*) is not adequately resolved by SMOTE or oversampling alone and constitutes a significant open challenge. The Spanish track does not suffer from this limitation.

Future work should explore LLM-based data augmentation for minority classes in English Subtask 3, cross-lingual transfer from the more balanced Spanish annotations, and the integration of richer sentiment and pragmatic features as auxiliary inputs to improve performance on group-targeted support detection.

Acknowledgments

The authors thank the organizers of the SSD-2026 Social Support Detection shared task [27] at IberLEF 2026 [28] for providing the dataset, annotation guidelines, and evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. Further, the authors used Claude in order to: Citation management, Formatting assistance. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Statista (2024). Social media and user-generated content — overview. Retrieved March 2026 from https://www.statista.com/markets/424/topic/540/social-media-user-generated-content/

[2] Mustofa, B. A., & Saptomo, W. L. Y. (2025). Use of natural language processing in social media text analysis. *Journal of Artificial Intelligence and Engineering Applications*, 4(2).

[3] Powell, J., et al. (2024). Evaluating large language models for health-related text classification tasks with public social media data. *JAMIA Open*.

[4] Hou, Y., & Huang, J. (2025). Natural language processing for social science research: A comprehensive review. *Journal of Digital Social Research*.

[5] Henning, P., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2023). A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of EACL 2023*, pp. 588–609.

[6] Garg, M., et al. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*.

[7] Hasan, K., Saquer, J., & Zhang, Y. (2025). Mental multi-class classification on social media: Benchmarking transformer architectures against LSTM models. In *2025 IEEE International Conference on Machine Learning and Applications (ICMLA)*. arXiv:2509.16542.

[8] Kolesnikova, O., Tash, M. S., Ahani, Z., Agrawal, A., Monroy, R., & Sidorov, G. (2025). Advanced machine learning techniques for social support detection on social media. *Heliyon*, 11(10). https://doi.org/10.1016/j.heliyon.2025.e43437

[9] Ahani, Z., Tash, M. S., Balouchzahi, F., Ramos, L., Sidorov, G., & Gelbukh, A. (2026). Social support detection from social media texts. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0337476

[10] Tash, M. S., Ramos, L., Ahani, Z., Monroy, R., Calvo, H., & Sidorov, G. (2025). Online social support detection in Spanish social media texts. *arXiv preprint arXiv:2502.09640*.

[11] Kirk, H., et al. (2023). SemEval-2023 Task 10: Explainable detection of online sexism. In *Proceedings of SemEval-2023*, pp. 2193–2210.

[12] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pp. 1532–1543.

[13] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*, pp. 1746–1751.

[14] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

[15] Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

[17] Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pp. 8440–8451.

[18] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of KDD 2016*, pp. 785–794.

[19] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

[20] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[21] Wolf, T., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020 (System Demonstrations)*, pp. 38–45.

[22] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1–5.

[23] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM 2014*.

[24] Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Nielsen, M. E. (2021). pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks. *arXiv preprint arXiv:2106.09462*.

[25] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

[26] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

[27] Shahiki Tash, M., Ramos, L., Zahra, A., Barrera Animas, A. Y., Balouchzahi, F., Kolesnikova, O., Sidorov, G., Gelbukh, A., Monroy, R., & Vargas, F. (2026). Overview of SSD-2026 at IberLEF 2026: Social Support Detection. 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.

[28] Bonet-Jover, A., González-Barba, J. A., & Chiruzzo, L. (2026). Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026)*, CEUR-WS.org.

[29] Shahiki-Tash, M., Ahani, Z., Ramos, L., et al. Linguistic Analysis in Different Types of Support using LIWC for Spanish and English online communities. https://doi.org/10.21203/rs.3.rs-9557287/v1

A. Supplementary Tables

Table 6: 5-fold CV results, Subtask 1 — English. Macro-F1, Accuracy, Macro-Precision, Macro-Recall, Weighted-F1. Best Macro-F1 in bold. Tier 2 aggregate metrics available only for the representative model shown.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 10$)	0.7878 ± 0.0051	0.8476	0.7752	0.7941	0.8460
XGBoost (depth 8, lr 0.2)	0.7742 ± 0.0150	0.8428	0.7561	0.7572	0.8296
Random Forest (depth unlimited)	0.7649 ± 0.0127	0.8579	0.8031	0.7166	0.8326
SVM ($C = 5$)	0.7573 ± 0.0154	0.8502	0.7865	0.7227	0.8310
Tier 2: Neural models with GloVe/FastText					
TextCNN [GloVe Twitter 200d]	0.7870 ± 0.0058	0.8460	0.7819	0.8084	0.8519
TextCNN [GloVe Twitter 100d]	0.7827 ± 0.0031	0.8415	0.7666	0.7986	0.8410
TextCNN [FastText Wiki-EN 100d]	0.7811 ± 0.0051	0.8402	0.7665	0.8097	0.8413
BiLSTM+Attn [FastText Wiki-EN 100d]	0.7690 ± 0.0117	0.8229	0.7588	0.8052	0.8352
MLP [GloVe Twitter 100d]	0.7601 ± 0.0064	0.8092	0.7403	0.8097	0.8149
Tier 3: Pre-trained transformer					
RoBERTa	0.7997 ± 0.0079	0.8516	0.7737	0.8240	0.8469

Table 7: 5-fold CV results, Subtask 2 — English. Metrics as in Table 6. Best Macro-F1 in bold.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 10$)	0.7597 ± 0.0305	0.8589	0.7622	0.7830	0.8587
XGBoost (depth 8, lr 0.2)	0.7192 ± 0.0288	0.8251	0.7504	0.7836	0.8522
Random Forest (depth unlimited)	0.7064 ± 0.0406	0.8584	0.8483	0.6941	0.8551
SVM ($C = 10$)	0.6846 ± 0.0401	0.8527	0.8230	0.7041	0.8556
Tier 2: Neural models with GloVe/FastText					
MLP [FastText Wiki-EN 100d]	0.8109 ± 0.0328	0.8810	0.8033	0.8103	0.8820
TextCNN [FastText Wiki-EN 100d]	0.8095 ± 0.0209	0.8855	0.8215	0.7904	0.8841
TextCNN [GloVe Twitter 200d]	0.8070 ± 0.0224	0.8883	0.8231	0.7711	0.8791
BiLSTM+Attn [GloVe Twitter 200d]	0.7955 ± 0.0226	0.8708	0.7672	0.8081	0.8638
Tier 3: Pre-trained transformer					
RoBERTa	0.8288 ± 0.0357	0.8996	0.8094	0.7677	0.8741

Table 8: 5-fold CV results, Subtask 3 — English. Metrics as in Table 6. Best Macro-F1 in bold.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 5$)	0.7303 ± 0.0537	0.8241	0.8136	0.8489	0.8381
XGBoost (depth 6, lr 0.1)	0.7149 ± 0.0607	0.8463	0.7565	0.7489	0.8717
Random Forest (depth unlimited)	0.7093 ± 0.1020	0.8373	0.9165	0.8071	0.8620
SVM ($C = 5$)	0.6480 ± 0.0413	0.8186	0.7556	0.6789	0.8370
Tier 2: Neural models with GloVe/FastText					
TextCNN [FastText Wiki-EN 100d]	0.7537 ± 0.0761	0.8456	0.8075	0.7462	0.8619
BiLSTM+Attn [GloVe Twitter 200d]	0.7353 ± 0.0683	0.8359	0.8161	0.8190	0.8665
MLP [FastText Wiki-EN 100d]	0.7075 ± 0.0891	0.8110	0.8015	0.8550	0.8527
Tier 3: Pre-trained transformer					
RoBERTa	0.7718 ± 0.0413	0.8629	0.7988	0.8449	0.8944

Table 9: 5-fold CV results, Subtask 1 — Spanish. Metrics as in Table 6. Best Macro-F1 in bold.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 10$)	0.8181 ± 0.0130	0.8704	0.8260	0.8318	0.8794
XGBoost (depth 6, lr 0.1)	0.8053 ± 0.0201	0.8646	0.8115	0.7962	0.8637
Random Forest (depth unlimited)	0.8011 ± 0.0183	0.8777	0.8492	0.7538	0.8590
SVM ($C = 5$)	0.7967 ± 0.0213	0.8673	0.8313	0.7820	0.8660
Tier 2: Neural models with FastText CC-ES					
TextCNN [FastText CC-ES 100d]	0.8062 ± 0.0158	0.8654	0.8091	0.7920	0.8615
BiLSTM+Attn [FastText CC-ES 100d]	0.7969 ± 0.0155	0.8581	0.8007	0.8147	0.8631
MLP [FastText CC-ES 100d]	0.7851 ± 0.0260	0.8381	0.8070	0.8363	0.8703
Tier 3: Pre-trained transformers					
XLM-R	0.8430 ± 0.0079	0.8850	0.8246	0.8378	0.8801
BETO	0.8395 ± 0.0085	0.8873	0.8323	0.8343	0.8829

Table 10: 5-fold CV results, Subtask 2 — Spanish. Metrics as in Table 6. Best Macro-F1 in bold.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 10$)	0.7260 ± 0.0478	0.8085	0.6702	0.6563	0.7663
Random Forest (depth unlimited)	0.7025 ± 0.0252	0.8102	0.7505	0.7153	0.8153
SVM ($C = 1$)	0.6978 ± 0.0352	0.8085	0.6667	0.6433	0.7626
XGBoost (depth 4, lr 0.05)	0.6259 ± 0.0461	0.8068	0.6769	0.5706	0.7345
Tier 2: Neural models with FastText CC-ES					
TextCNN [FastText CC-ES 100d]	0.8026 ± 0.0422	0.8746	0.8654	0.7298	0.8492
MLP [FastText CC-ES 100d]	0.7895 ± 0.0296	0.8610	0.7784	0.7393	0.8329
BiLSTM+Attn [FastText CC-ES 100d]	0.7827 ± 0.0336	0.8458	0.8439	0.8059	0.8779
Tier 3: Pre-trained transformers					
XLM-R	0.8617 ± 0.0445	0.9000	0.8417	0.8504	0.8905
BETO	0.8305 ± 0.0399	0.8780	0.7824	0.8099	0.8510

Table 11: 5-fold CV results, Subtask 3 — Spanish. Metrics as in Table 6. Best Macro-F1 in bold.

Model	Macro-F1	Acc	M-Prec	M-Rec	W-F1
Tier 1: Sparse traditional models					
Logistic Regression ($C = 10$)	0.7392 ± 0.0343	0.8041	0.8048	0.6844	0.7804
XGBoost (depth 4, lr 0.05)	0.7191 ± 0.0658	0.7773	0.7466	0.6110	0.7065
Random Forest (depth unlimited)	0.7140 ± 0.0297	0.7711	0.8178	0.6857	0.7648
SVM ($C = 10$)	0.6975 ± 0.0522	0.7711	0.7418	0.5949	0.7132
Tier 2: Neural models with FastText CC-ES					
MLP [FastText CC-ES 100d]	0.7090 ± 0.0765	0.7876	0.6689	0.6790	0.7576
TextCNN [FastText CC-ES 100d]	0.6682 ± 0.0212	0.7732	0.7485	0.6234	0.7273
BiLSTM+Attn [FastText CC-ES 100d]	0.6580 ± 0.0717	0.7588	0.6702	0.6891	0.7646
Tier 3: Pre-trained transformers					
XLM-R	0.8585 ± 0.0441	0.8784	0.8436	0.8759	0.8541
BETO	0.8380 ± 0.0412	0.8825	0.8716	0.8500	0.8843

B. Supplementary Figures

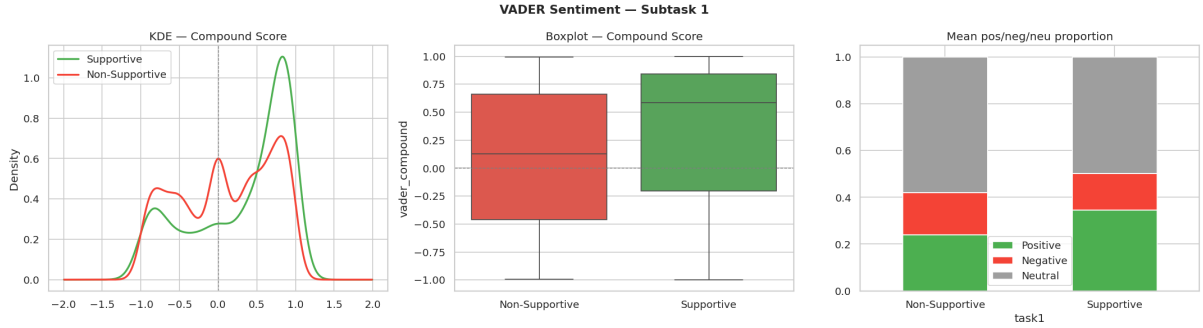

Figure 3: VADER compound sentiment distribution by class, Subtask 1 — English. Supportive comments show a strong concentration of positive scores; non-supportive comments are more broadly distributed.

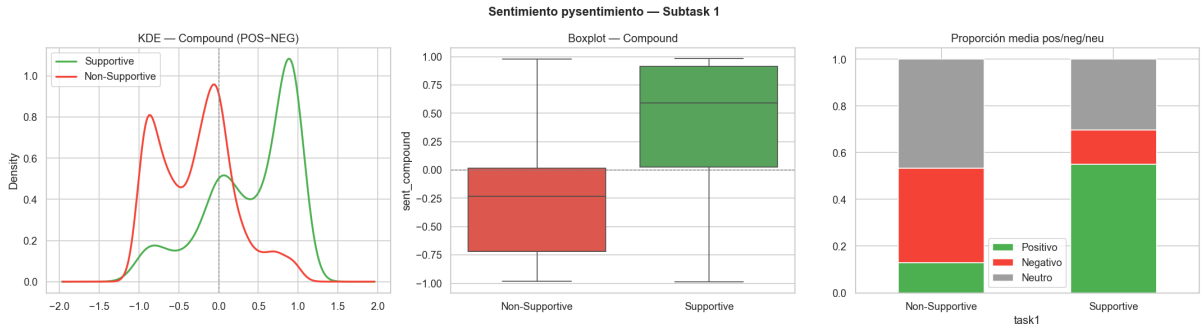

Figure 4: Pysentimiento compound sentiment distribution by class, Subtask 1 — Spanish. The separation between Supportive ($+0.405$) and Non-Supportive (-0.277) is more pronounced than in the English corpus.

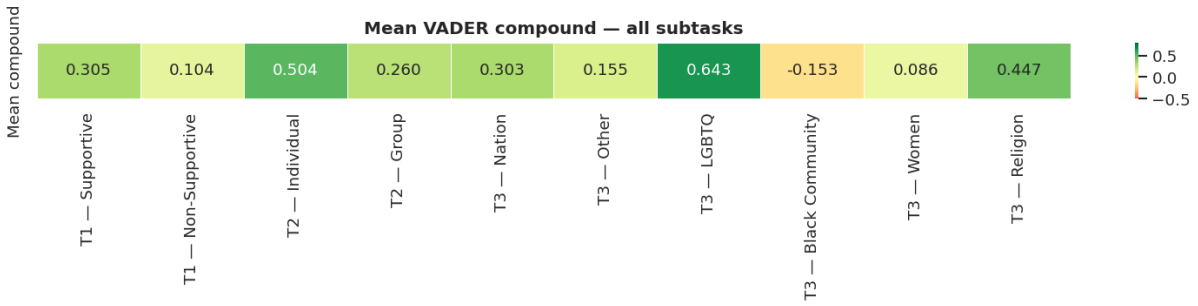

Figure 5: Mean VADER compound score across all subtask classes — English. The highest score is for the LGBTQ group (T3, +0.643); the only negative mean is for Black Community (−0.153), reflecting the more conflictual discourse in that class.

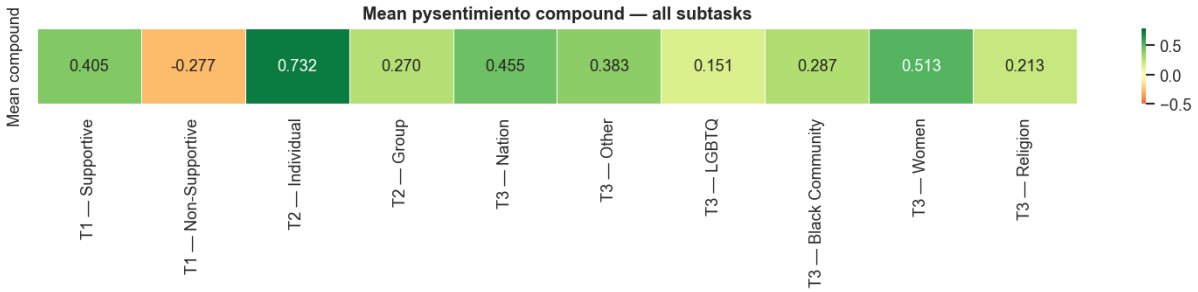

Figure 6: Mean pysentimiento compound score across all subtask classes — Spanish. The highest score is for Individual support (T2, +0.732); Non-Supportive comments have the only negative mean (−0.277), confirming the strong sentiment contrast at the Subtask 1 level.

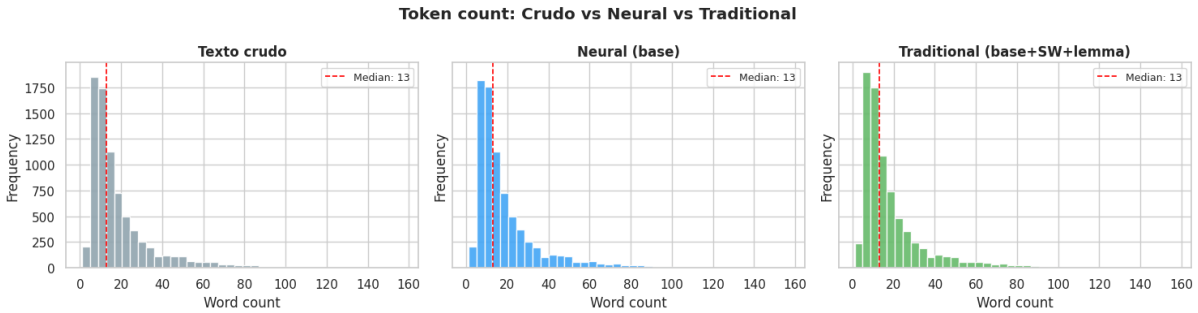

Figure 7: Token count distribution across preprocessing stages — English. Raw text (median 13), neural base pipeline (median 13), and traditional pipeline with stopwords and lemmatisation (median 13). The distribution shape is preserved across all stages.

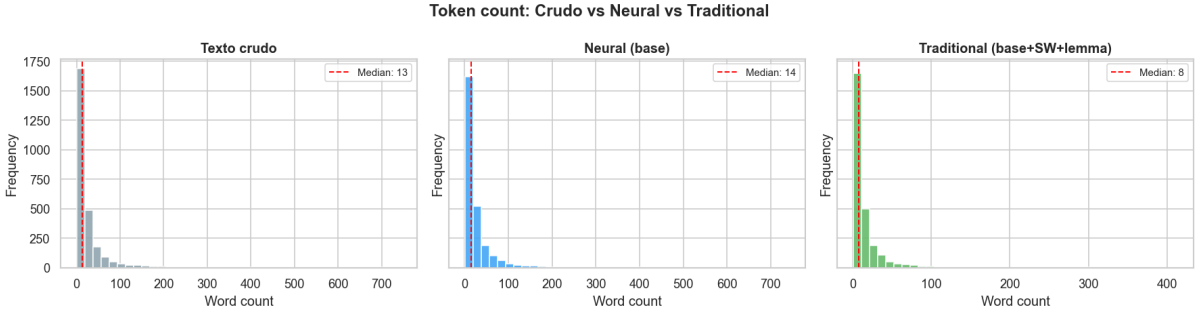

Figure 8: Token count distribution across preprocessing stages — Spanish. Raw text (median 13), neural base pipeline with emoji mapping (median 14), and traditional pipeline (median 8). The Spanish traditional pipeline achieves stronger length reduction due to Spanish morphological richness.