Comparing Transformer Fine-Tuning, Multi-Task Learning, and LLM-as-Judge for Social Support Detection

Leonardo Kenji Minemura Suazo[1], José Alexander Romero Mitiaeva[1]

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

Abstract

Social support detection — identifying encouragement, solidarity, and care in social media text — requires fine-grained emotion and sentiment analysis beyond standard polarity classification. We participate in SSD-2026 at IberLEF 2026, a hierarchical shared task using 10,598 released training comments and, after label release, 2,651 official test comments: Subtask 1 detects whether a comment is supportive; Subtask 2 identifies whether support targets an individual or a group; and Subtask 3 classifies the targeted community (Nation, LGBTQ, Black Community, Women, Religion, or Other). We compare five approaches: TF-IDF baselines, FastText deep learning models, domain-adapted transformer fine-tuning (BERTweet, RoBERTuito, XLM-R), multi-task learning (XLM-Twitter), and LLM-as-judge prompting with GPT-5.4-mini using emotion-aware few-shot examples. Our best submissions achieve official CodaBench scores of 0.8231 for English and 0.8785 for Spanish. After the official labels were released, retrospective 10-fold BERTweet and RoBERTuito ensembles achieved average per-subtask official-test macro F1 of 0.829 and 0.890, respectively.

Keywords

social support detection, transformer fine-tuning, multi-task learning, LLM-as-judge, BERTweet, cross-lingual NLP, shared task

CEUR-WS.org/Vol-4273/SSD-2026_paper12.pdf

1. Introduction

Natural language processing (NLP) has driven rapid progress in understanding human language at scale, enabling automated analysis of text that would be infeasible by manual inspection alone [1, 2]. Social media platforms — Twitter, Reddit, YouTube, Instagram — have become primary channels for public discourse, generating billions of messages daily that reflect opinions, emotions, and interpersonal dynamics in real time [3, 4]. This volume of user-generated text presents both an opportunity and a challenge for NLP: while the data is abundant, it is noisy, informal, and context-dependent.

Within this landscape, the intersection of social support and social media has received growing attention from both the social sciences and the NLP community [5, 6]. Social support — acts of encouragement, solidarity, care, or help directed at others — plays a critical role in mental health, community resilience, and crisis response [7, 8]. Detecting such supportive language in social media enables applications ranging from monitoring community wellbeing to identifying vulnerable populations in need of intervention [9, 10]. However, this task is distinct from sentiment analysis: a comment can be emotionally neutral yet deeply supportive, emotionally charged yet not supportive of anyone, and the *target* of support (an individual versus a community) matters for downstream applications.

The gap. Despite this importance, most NLP research on social media text has focused on detecting *negative* phenomena — hate speech, toxicity, offensive language [11, 12] — while *prosocial* language remains comparatively understudied. Similarly, sentiment analysis often reduces affect to positive or negative polarity, whereas supportive sentiment requires detecting directed care, encouragement, and solidarity. Existing work on social support detection has been limited in scope: early studies relied on lexicon-based and feature-engineering methods applied to specific health forums [8], and more recent transformer approaches have been evaluated on social support datasets [13, 14]. Key open questions remain: How do domain-adapted transformer models compare to general-purpose multilingual mod-

IberLEF 2026, September 2026, León, Spain

EMAIL: 0246757@up.edu.mx (L. K. M. Suazo); 0248744@up.edu.mx (J. A. R. Mitiaeva)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-09-29

els and large language models (LLMs) for this task? Does multilingual training help in data-scarce subtasks? Can LLMs complement fine-tuned classifiers, particularly for underrepresented classes?

The shared task. The Social Support Detection shared task at IberLEF 2026 (SSD-2026) [15, 16] addresses these gaps by formalizing social support detection as a three-level hierarchy: Subtask 1 determines whether a comment is supportive; Subtask 2 classifies support as targeting an individual or a group; and Subtask 3 identifies the specific community (Nation, Other, LGBTQ, Black Community, Religion, or Women). The task covers both English (7,998 training samples) and Spanish (2,600 samples) social media text, with pronounced class imbalance — as few as 15 training examples for the Religion class. The shared-task setup and evaluation context are described in the SSD-2026 overview paper and the IberLEF 2026 overview paper [16, 15]. The dataset and annotation framework build on prior English social support detection work [13], transformer benchmarking for social support detection [14], Spanish social support detection [17], and LIWC-based analyses of support types [18].

Our contribution.

- We systematically compare five modeling paradigms for social support detection across English and Spanish: TF-IDF baselines, FastText neural models, single-task transformers, multi-task learning, and LLM-as-judge prompting.
- We frame social support detection as emotion and sentiment analysis of prosocial language, using emotion-aware prompting to separate directed support from undirected emotional reactions.
- We show that multilingual training improves English minority-class performance, with XLM-R improving Subtask 3 by 12.5 F1 points over English-only training.
- We find that hybrid LLM+transformer verification is limited by anchoring bias: the LLM over-agrees with provided pre-labels instead of independently correcting errors.
- Our official CodaBench scores are 0.8231 for English (three-way tie) and 0.8785 for Spanish.
- We add a post-release official-label analysis using stratified 10-fold cross-validation and official-test ensemble evaluation, while keeping the CodaBench leaderboard scores as the shared-task competition result.

2. Related Work

Social support detection. Early NLP work on social support drew from social science frameworks distinguishing informational, emotional, and instrumental support [7], applying lexicon-based and feature-engineering methods to health forums and Reddit communities [8, 9]. Neural approaches later showed that deep models capture subtle supportive cues that surface-level features miss [10]. Most directly related to our work, Ahani et al. [13] formalize social support detection for social media texts, and Kolesnikova et al. [14] benchmark transformer approaches on this setup, with RoBERTa [19] reaching macro F1 of 0.78–0.84 across subtasks. Tash et al. [17] extend the framework to Spanish social media, demonstrating cross-lingual applicability. Adjacent work on hope speech [20] and empathetic dialogue [21] further shows that positively directed prosocial language is tractable with transformer-based classifiers [22].

Transformer models for social media. BERT [2] and its variants have become the standard backbone for social media classification. For cross-lingual scenarios, Conneau et al. [23] introduced XLM-R, a RoBERTa model pretrained on 2.5TB of filtered Common Crawl data across 100 languages, which achieves strong zero-shot cross-lingual transfer. Domain-adapted models offer additional gains on informal text: Nguyen et al. [24] pretrained BERTweet on 850 million English tweets, and Pérez et al. [25] released RoBERTuito on over 500 million Spanish tweets, both outperforming general-purpose models on Twitter classification tasks [3, 4].

Target identification and class imbalance. The hierarchical structure of SSD-2026 — detect, then classify the target — mirrors shared tasks on offensive language such as OffensEval [11, 26] and HatEval [12], which require identifying whether content targets an individual or a group across English and Spanish Twitter. Work on community-targeted classification [27] has shown that minority groups are consistently harder to classify, a finding directly applicable to Subtask 3 where Women (19 examples) and Religion (15 examples) are severely underrepresented. We address this through inverse-frequency class weighting [14], a lighter-weight alternative to oversampling [28] and focal loss [29].

3. Task and Dataset Description

3.1. Task Definition

SSD-2026 is organized as a hierarchical three-subtask shared task on English and Spanish social media comments. A comment that expresses encouragement, care, admiration, help, or solidarity toward someone is labeled as *Supportive*; all others, including comments that promote violence or harm even if superficially positive, are labeled *Not Supportive*.

Subtask 1 — Support Detection (Binary). Given any social media comment, predict whether it is Supportive or Not Supportive.

Subtask 2 — Target Type Identification. Given a comment already labeled Supportive, predict whether the support targets an *Individual* or a *Group*.

Subtask 3 — Targeted Group Classification. Given a Supportive comment that targets a group, predict the specific community from six classes: Nation, Other, LGBTQ, Black Community, Religion, and Women.

3.2. Dataset

The official training set consists of 7,998 English-language and 2,600 Spanish-language social media comments drawn from diverse platforms. The official test set consists of 2,000 English-language and 651 Spanish-language comments; labels for these comments were unavailable during the competition phase and were used only for retrospective post-release analysis. The dataset and annotation framework build on prior work on English and Spanish social support detection and linguistic analysis of support types [13, 14, 17, 18]. Table 1 summarizes the label distributions for both languages across all three subtasks.

The dataset exhibits pronounced class imbalance throughout. For Subtask 1, only 22.4% of English comments are labeled Supportive (1,795 out of 7,998), while Spanish shows a similar ratio at 22.7% (590 out of 2,600), motivating the use of class-weighted loss during training. Subtask 2 is applied exclusively to Supportive comments: in English, the majority target a group (1,450) rather than an individual (345), yielding a roughly 4:1 imbalance; Spanish shows a comparable 3.4:1 ratio (456 vs. 134). Subtask 3 further partitions the group-supportive comments into six community categories. In English, Nation (786) and Other (416) dominate while Women (19) and Religion (15) are severely underrepresented. The Spanish distribution is strikingly different: LGBTQ (196) is the largest class, while Nation (47) and Black Community (24) are the rarest, reflecting distinct social media discourse contexts across languages.

Representative training cases illustrate why the task goes beyond polarity: emotional reactions, political commentary, and directed support can use overlapping vocabulary while requiring different labels. For example, solidarity phrases such as “free palestine” may be Supportive and Group-targeted, while emotionally intense commentary without directed support remains Non-Supportive.

We use the label distributions in Table 1, together with text-length and lexical checks, to guide preprocessing and class weighting. Supportive comments are slightly longer on average than Non-Supportive comments in both languages, but the distributions are heavily right-skewed and overlap substantially. English bigram and trigram checks show that supportive and non-supportive comments

Table 1
Label distribution across all subtasks for English and Spanish training data.

Subtask	Label	English	Spanish
1 — Support	Non-Supportive	6,203	2,010
	Supportive	1,795	590
2 — Target	Group	1,450	456
	Individual	345	134
3 — Community	Nation	786	47
	Other	416	81
	LGBTQ	123	196
	Black Community	91	24
	Women	19	48
	Religion	15	60

often discuss the same topics; the distinction depends less on keywords alone than on whether encouragement, care, or solidarity is directed at a person or group.

4. Methodology

4.1. Pipeline and Evaluation Protocol

We treat SSD-2026 as a hierarchical inference problem. At prediction time, Subtask 1 is applied to every comment; Subtask 2 is applied only to comments predicted as Supportive; and Subtask 3 is applied only to comments predicted as both Supportive and Group-targeted. This sequential design matches the shared task definition and prevents irrelevant Subtask 2 and Subtask 3 labels from being assigned to Non-Supportive comments.

All internal experiments are evaluated with macro-averaged F1, the official shared task metric, computed separately for each subtask. For the original shared-task submission, model selection used the best validation macro F1 from the training-set split and we report the best single seed together with the mean and standard deviation over three seeds where available. During this phase, the official test labels were hidden; submitted systems were trained and selected only from the released training data, and CodaBench provided the only feedback on the held-out test set. Official leaderboard results are reported separately in Section 6 because CodaBench scores aggregate the official test performance for a language submission and are not interchangeable with per-subtask internal validation F1. After the official test labels were released, we additionally ran a post-release evaluation protocol: stratified 10-fold cross-validation on the training data for the strongest language-specific transformers (BERTweet for English and RoBERTuito for Spanish) and the TF-IDF baselines, followed by official-test inference with 10-fold model ensembles. These retrospective experiments are used for robustness analysis and error analysis, but they do not replace the official CodaBench leaderboard results.

4.2. Data Splits

The full training set is partitioned into train, validation, and test subsets at an 80/10/10 ratio using stratified sampling. For Subtask 2, only Supportive-labeled examples are retained, yielding approximately 1,436 / 179 / 180 English examples. For Subtask 3, only Group-targeted examples are retained, reducing to approximately 1,160 English training examples across six categories. The same stratified procedure is used for Spanish and multilingual experiments. For the post-release experiments, we use 10-fold cross-validation, a standard choice for accuracy estimation and model selection in supervised learning [30]. Folds are generated at the comment level on the complete training set before subtask-specific filtering. This preserves the hierarchical task structure and prevents the same comment from appearing in both training and validation partitions across subtasks.

4.3. Training Configuration

We keep the training setup consistent across comparable models. Traditional baselines use combined word and character TF-IDF features (up to 50,000 features) with lightweight handcrafted features. TextCNN and BiLSTM models use 300-dimensional FastText embeddings, learning rate 10^{-3}, batch size 64, dropout 0.3, maximum length 128, and early stopping. Fine-tuned transformers use AdamW, learning rate 2×10^{-5}, batch size 32, maximum length 128, weight decay 0.01, 10% warmup, gradient clipping at ℓ_2 norm 1.0, inverse-frequency class-weighted cross-entropy, and three random seeds (13, 42, 77). The multi-task XLM-Twitter model first trains only task heads (10^{-4}, 3 epochs) and then fine-tunes the shared encoder (2×10^{-5}) with equal task weights. LLM-as-judge experiments use GPT-5.4-mini at temperature 0.0 with 26 few-shot examples.

5. System Description

5.1. Overall Approach

We explore five modeling paradigms for each subtask:

1. **Traditional ML**: Logistic Regression and SVM with TF-IDF features augmented by handcrafted features (text length, punctuation counts, emoji density).
2. **Deep Learning**: TextCNN and BiLSTM with pretrained FastText embeddings.
3. **Single-Task Transformers**: BERTweet (English), RoBERTuito (Spanish), and XLM-R (both languages plus multilingual training).
4. **Multi-Task Learning**: A shared XLM-Twitter encoder with three task-specific classification heads, trained jointly on all subtasks.
5. **LLM-as-Judge**: GPT-5.4-mini with a carefully engineered combined few-shot prompt classifying all three subtasks simultaneously.

For the competition submission, we adopt a *pipeline* strategy: three independent models are applied sequentially, each consuming only the subset of examples relevant to its subtask. The best model for each subtask is selected based on validation macro F1.

5.2. Text Preprocessing

We apply different preprocessing pipelines depending on the model family:

- **Traditional ML**: Aggressive preprocessing — lowercasing, URL and mention removal, emoji-to-text conversion, slang expansion, and lemmatization with spaCy.
- **Deep Learning**: Moderate preprocessing — lowercasing, URL and mention removal, emoji-to-text conversion.
- **Transformers and LLM**: Minimal preprocessing — Unicode normalization (NFKC), emoji-to-text conversion via the `emoji` library, and whitespace collapsing. We let the subword tokenizer handle the rest, preserving casing and punctuation that carry semantic signal in social media.

5.3. Traditional ML Baselines

We train Logistic Regression (LR) and linear SVM classifiers on combined TF-IDF features (word unigrams + character 3–5-grams, max 50,000 features) augmented with handcrafted features including text length, exclamation mark count, question mark count, and uppercase ratio.

5.4. Deep Learning Models

We experiment with TextCNN [3] using filter sizes [2, 3, 4, 5] and BiLSTM with 256-dimensional hidden states, both initialized with 300-dimensional FastText embeddings. Both models use dropout (0.3), class-weighted cross-entropy loss, and early stopping on validation macro F1 with patience of 5–7 epochs.

5.5. Single-Task Transformer Fine-Tuning

Each subtask model fine-tunes a pretrained transformer encoder with a linear classification head over the `[CLS]` token. We evaluate three encoders:

- **BERTweet** [24]: Pretrained on 850M English tweets; used for English subtasks.
- **RoBERTuito** [25]: Pretrained on 500M+ Spanish tweets; used for Spanish subtasks.
- **XLM-R** [23]: Multilingual RoBERTa pretrained on 100 languages; evaluated on English-only, Spanish-only, and combined multilingual training data.

The shared training settings, including maximum sequence length, class weighting, early stopping, and multi-seed evaluation, are detailed in Section 4.

5.6. Multi-Task Learning

We train a multi-task model with a shared XLM-Twitter (`cardiffnlp/twitter-xlm-roberta-base`) encoder and three task-specific linear classification heads. Training proceeds in two phases: (1) a warm-up phase (3 epochs) training only the classification heads with learning rate 10^{-4}, followed by (2) end-to-end fine-tuning of all parameters with learning rate 2×10^{-5}. The total loss is a weighted sum of the three task-specific cross-entropy losses with equal weights ($\lambda_1 = \lambda_2 = \lambda_3 = 1$). Gradient accumulation (2 steps) with batch size 16 yields an effective batch size of 32.

5.7. LLM-as-Judge

We use GPT-5.4-mini as a zero-resource classifier via the OpenAI API. A single combined prompt instructs the model to classify all three subtasks at once, returning a JSON object with keys `task1`, `task2`, and `task3`. The prompt includes:

- Detailed task definitions with explicit boundary conditions (e.g., "admiration directed at a person is Supportive; general praise of qualities is Non-Supportive").
- 26 few-shot examples drawn from the training data, including deliberately tricky cases (emotional reactions to tragedy, sports fan debate, religious preaching, political opposition).
- Temperature set to 0.0 for deterministic output.

5.8. LLM Prompt Design

The final prompt combines a system role, task definitions, emotion-aware boundary rules, 26 few-shot examples, and a strict JSON output schema. The key design decision was to distinguish undirected emotional reactions from directed support: comments such as "tears come to my eyes" express emotion but are Non-Supportive unless they include encouragement, care, or solidarity toward a person or group.

The prompt asks the model to classify all three subtasks simultaneously and return a strict JSON object. It defines Supportive comments as encouragement, solidarity, care, help, or admiration directed at a person or group, while analytical statements, harmful text, and emotional reactions without directed support are Non-Supportive. The prompt was refined over eight versions on a 100-sample development subset; the most useful revisions tightened the Non-Supportive boundary and added the annotation rule that admiration directed at a person can be Supportive. The final prompt reached 0.833 Subtask 1 macro F1 on a separate 200-sample internal subset. These LLM subset results are used only as diagnostic prompt-development evidence and are not directly comparable with validation or official-test scores.

Hybrid: BERTweet + LLM verification. We additionally experiment with a hybrid approach where BERTweet predictions are provided to the LLM as "pre-labels" alongside the original text. The LLM is instructed to either confirm or correct the BERTweet prediction. We find that this approach suffers from anchoring bias: the LLM agrees with the pre-label 88–97% of the time depending on the subtask, limiting its ability to correct errors.

6. Results

We report four evaluation regimes. First, internal validation results are computed from stratified 80/10/10 splits of the official training data. Baseline runs use seed 42; transformer experiments use seeds 13, 42, and 77, with best-seed and mean/std results reported where available. Second, LLM prompt-development results are measured on 100- and 200-sample internal subsets due to API cost; these subset results are diagnostic and are not directly comparable with validation, CodaBench, or post-release official test results. Third, official CodaBench scores are leaderboard scores on the hidden test set and are not directly comparable to internal per-subtask F1. Fourth, after the official labels were released, we evaluated archived submissions and cross-validation ensembles against the labeled official test set. We treat these post-release numbers as retrospective analysis, not as the competition ranking.

6.1. Official CodaBench Results

Table 2 reports the official CodaBench scores from the SSD-2026 leaderboard. These scores are macro-averaged leaderboard scores for a complete language submission and should be distinguished from the internal per-subtask validation results in Table 4.

Table 2
Official CodaBench scores for submitted systems.

Submission	Lang.	Strategy	Score
submission_english.zip	EN	Initial English pipeline	**0.8231**
submission_english_v2.zip	EN	English pipeline v2	**0.8231**
submission_english_bertweet.zip	EN	BERTweet-only pipeline	**0.8231**
submission_english_full.zip	EN	Full pipeline	0.8201
submission_en_hybrid_bertweet_v2.zip	EN	BERTweet+LLM hybrid	0.8016
submission_english_llm.zip	EN	LLM judge pipeline	0.7918
submission_spanish_full.zip	ES	Full pipeline	**0.8785**
submission_spanish.zip	ES	RoBERTuito-only pipeline	0.8544
submission_spanish_v3.zip	ES	Spanish v3 pipeline	0.8459
submission_es_hybrid_v3.zip	ES	Spanish hybrid pipeline	0.7935

Three English submissions tied at the best official English score of 0.8231. Our full English pipeline scored 0.8201, marginally below the tied best; adding XLM-R multilingual for Subtask 3 improved internal minority-class behavior but did not improve the aggregate official score, possibly because it over-predicted minority classes on the official test distribution. The Spanish full pipeline achieved the best Spanish score, 0.8785.

6.2. Post-release Official-label Evaluation

After the official labels were released, we evaluated both the submitted prediction files and 10-fold cross-validation ensembles directly against the official test labels. This analysis has a different purpose from Table 2: the CodaBench table preserves the competition record, while Table 3 reports per-subtask diagnostic performance using labels that were unavailable during the shared-task phase. The released

official test labels are used only for this retrospective evaluation; all post-release cross-validation models are still trained only on the official training data.

The local re-evaluation of archived submissions confirms the leaderboard ordering for Subtask 1. Averaging the three subtasks locally gives 0.8372 for the English full pipeline and 0.8843 for the Spanish full pipeline; these local averages are post-release diagnostics, not replacement leaderboard scores.

Table 3
Post-release 10-fold cross-validation and official-test ensemble results for the selected language-specific transformers. CV values are mean macro-F1 ± standard deviation over folds; official-test values are macro-F1 from the 10-fold ensemble.

		10-fold CV			Official test ensemble			
Lang.	**Model**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**Avg.**
EN	BERTweet	.813 ± .015	.855 ± .031	.758 ± .093	.823	.871	.793	.829
ES	RoBERTuito	.862 ± .019	.897 ± .056	.839 ± .076	.884	.925	.861	.890

The post-release analysis supports the original model choices. BERTweet remains strong for English, RoBERTuito remains strong for Spanish, and Subtask 3 shows the largest variance because its minority classes are sparse. The ensembles are therefore best read as robustness checks rather than new competition entries.

6.3. Validation Results

Table 4 summarizes the main validation results with macro-F1 for each subtask. Transformer means are reported where multi-seed runs were available. LLM rows come from smaller internal subsets and are included only for diagnostic comparison.

Table 4
Validation macro-F1 by language and subtask. Bold marks the best fine-tuned model per language/subtask; the LLM row is subset-based and not directly comparable with validation or official-test scores.

Lang.	Approach	Model	T1	T2	T3	Mean F1
EN	Trad. ML	LR + TF-IDF	.803	.775	.721	—
EN	Deep Learn.	TextCNN	.778	.829	.836	.776 ± .002
EN	Single-Task	BERTweet	**.831**	**.896**	.764	.826 ± .005
EN	Single-Task	XLM-R	.801	.836	.789	.792 ± .007
EN	XLM-R Multi	XLM-R	.829	.846	**.858**	.847 ± .009
EN	Multi-Task	XLM-T	.793	.903	.847	—
EN	LLM Judge	GPT-5.4-mini	.833	.857	.944	—
ES	Trad. ML	LR + TF-IDF	.827	.832	.822	—
ES	Single-Task	RoBERTuito	.865	**.953**	**.974**	.843 ± .022
ES	Single-Task	XLM-R	**.877**	.753	.736	.836 ± .031
ES	Multi-Task	XLM-T	.832	.823	.901	—

BERTweet is the strongest English fine-tuned model for Subtasks 1–2, while multilingual XLM-R is strongest for English Subtask 3, where Spanish examples help rare classes. The LLM judge performs well on its internal subset, especially for Subtask 3, but those results should be treated as indicative rather than evidence of general superiority. For Spanish, XLM-R obtains the best single-seed Subtask 1

result, while RoBERTuito is strongest and more stable for Subtasks 2–3; the final Spanish submission therefore combines XLM-R for Subtask 1 with RoBERTuito for Subtasks 2–3.

6.4. Cross-Lingual and Multilingual Analysis

Training XLM-R on combined English + Spanish data ("multilingual") consistently improves English results compared to English-only training: Subtask 1 rises from 0.792 to 0.815, Subtask 2 from 0.821 to 0.832, and Subtask 3 from 0.722 to 0.847. The largest improvement is on Subtask 3, where additional Spanish examples for rare classes partly alleviate English-only sparsity.

6.5. Hybrid Approach: BERTweet + LLM Verification

In the hybrid approach, BERTweet pre-labels are shown to the LLM with the instruction to confirm or correct them. On the official English test inputs (2,000 samples, unlabeled during the submission phase), the LLM changed 145 Subtask 1 labels (7.25%), 43 Subtask 2 labels, and 32 Subtask 3 labels. The high agreement rate (93–97%) indicates strong anchoring bias: the LLM tends to defer to the suggested label rather than independently evaluate. This limits the hybrid approach's ability to correct systematic errors in the pre-labels.

6.6. Error Analysis

Class imbalance effects. All fine-tuned models struggle with the Religion class (15 training examples). BERTweet achieves 0.0 F1 for Religion across most seeds, while XLM-R multilingual reaches 0.82 by leveraging Spanish Religion examples. The LLM judge also handles some Religion examples through in-context examples, but this evidence comes from a smaller internal subset and should be interpreted cautiously.

Figure 1 shows representative post-release official-test confusion matrices for the English BERTweet 10-fold ensemble. The dominant Subtask 1 error is missing Supportive comments, and Subtask 3 errors mainly fall between Nation and Other, with rare classes remaining unstable.

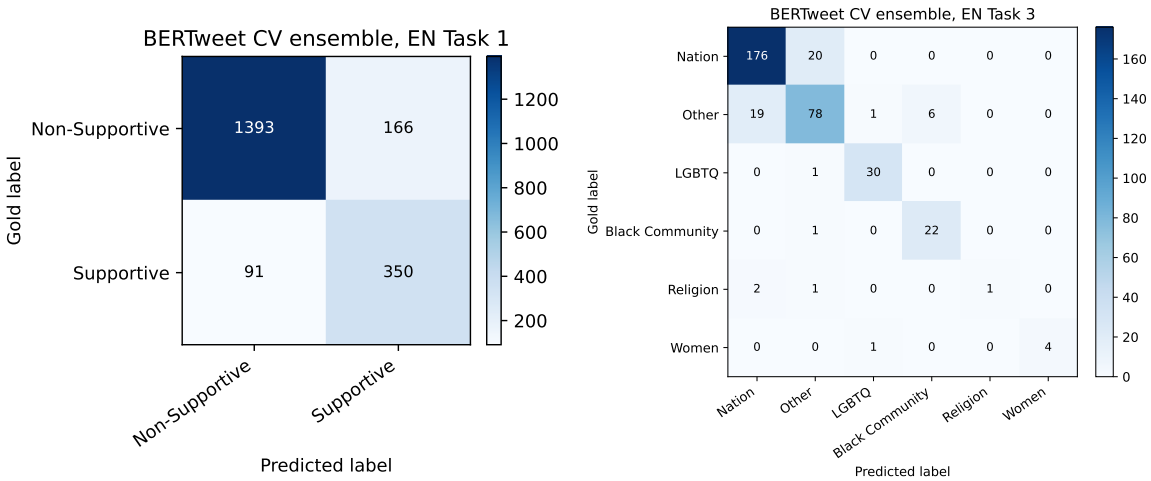

Figure 1: Official-test confusion matrices for the BERTweet 10-fold ensemble on English Subtasks 1 and 3.

In the corresponding Spanish official-test evaluation, Subtask 1 has fewer Supportive false negatives, and Subtask 3 remains strong despite the smaller training set, with most errors concentrated in the broader Nation, Other, and LGBTQ categories.

Boundary cases and seed sensitivity. The most challenging Subtask 1 cases distinguish directed support from emotional reactions, such as comments that express sadness about Palestine without

directly supporting a person or group. Transformer models also show seed sensitivity on smaller subtasks: BERTweet Subtask 2 ranges from 0.819 to 0.896 across seeds, and XLM-R ES Subtask 3 ranges from 0.276 to 0.736.

7. Future Work and Limitations

The main limitation of the LLM-as-judge experiments is evaluation scale. Internal prompt iterations were measured on 100–200-sample subsets because API cost made full repeated evaluation impractical, and the official LLM-only submission score (0.7918) is lower than the strongest internal subset results. The hybrid BERTweet+LLM strategy is also expensive while providing limited benefit because the LLM tends to anchor on the supplied pre-label. We also do not report statistical significance tests; differences between systems, especially on small Subtask 3 classes, should therefore be interpreted as empirical comparisons rather than confirmed significant effects.

The dataset itself imposes additional limits. Extreme class imbalance makes minority-class estimates unstable: in English, Religion has only 15 training examples and Women has 19, so a small number of errors can substantially shift macro F1. Spanish contains 2,600 training comments, roughly one third of the English training size, which limits robust cross-lingual generalization despite the strong performance of RoBERTuito. Because Subtask 3 assigns support to sensitive community categories such as LGBTQ, Religion, Women, Black Community, and Nation, deployment requires care. Misclassification may produce different reliability across communities, especially for rare classes, and systems of this kind could be misused for surveillance of political, religious, or identity-based discourse. We therefore view the submitted systems as shared-task classifiers rather than deployment-ready moderation or monitoring tools.

Future work should explore minority-class augmentation, larger multilingual encoders, and hybrid LLM designs that pass transformer confidence scores rather than explicit candidate labels, reducing anchoring bias. The emotion and sentiment signals identified here should also be tested on other languages and platforms.

8. Conclusion

We presented a comprehensive comparison of five modeling paradigms for the SSD-2026 social support detection task. Domain-adapted transformers are the most reliable systems: BERTweet is strong for English, RoBERTuito is strong for Spanish, and multilingual XLM-R improves English Subtask 3 by adding signal for rare classes. On the official CodaBench leaderboard, our best submissions achieved 0.8231 for English and 0.8785 for Spanish. Post-release official-label evaluation shows that 10-fold BERTweet and RoBERTuito ensembles achieve average per-subtask official-test macro F1 of 0.829 and 0.890, respectively. LLM-as-judge prompting is useful for analysis and may help diagnose minority-class behavior, but its subset-based results should not be treated as directly comparable to validation or official test scores. Hybrid LLM verification is limited by anchoring bias when the model is shown transformer pre-labels.

Acknowledgments

The authors thank the organizers of SSD-2026 for providing the dataset and evaluation framework.

Declaration on Generative AI

The authors used generative AI in two ways. First, GPT-5.4-mini was used as an experimental LLM-as-judge classifier through the OpenAI API, as described in the methodology and results sections. Second, OpenAI Codex was used during manuscript preparation for language editing, LaTeX maintenance, and

consistency checks. All generated outputs were reviewed, verified, and edited by the authors, who take full responsibility for the content of this paper.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2019).

[3] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, ACM, 2017, pp. 759–760.

[4] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, in: Computers in Human Behavior, volume 93, Elsevier, 2019, pp. 333–345.

[5] N. S. Coulson, Receiving social support online: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome, CyberPsychology & Behavior 8 (2005) 580–584.

[6] H. J. Oh, C. Lauckner, J. Boehmer, R. Fewins-Bliss, K. Li, Web-based social support and health: A meta-analysis, Computers in Human Behavior 38 (2014) 533–545.

[7] C. E. Cutrona, J. A. Suhr, Controllability of stressful events and satisfaction with spouse support behaviors, Communication Research 19 (1992) 154–174.

[8] W. Wang, L. Chen, K. Thirunarayan, A. P. Sheth, Harnessing twitter "big data" for automatic emotion identification, in: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing, IEEE, 2012, pp. 587–592.

[9] J. Zhang, Y. Luo, Y. Zhu, Understanding mental health support in online social networks, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2017, pp. 1–6.

[10] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 3838–3844.

[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 75–86.

[12] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.

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

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

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

[16] M. Shahiki Tash, L. Ramos, Z. Ahani, A. Y. Barrera Animas, F. Balouchzahi, O. Kolesnikova,

G. Sidorov, A. Gelbukh, R. Monroy, F. Vargas, 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, 2026.

[17] M. S. Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, G. Sidorov, Online social support detection in spanish social media texts, arXiv preprint arXiv:2502.09640 (2025). URL: https://arxiv.org/abs/2502.09640. arXiv:2502.09640.

[18] M. Shahiki-Tash, Z. Ahani, L. Ramos, et al., Linguistic analysis in different types of support using LIWC for spanish and english online communities, 2026. URL: https://doi.org/10.21203/rs.3.rs-9557287/v1. doi:10.21203/rs.3.rs-9557287/v1.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[20] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, 2020, pp. 41–53.

[21] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, Towards empathetic open-domain conversation models: A new benchmark and dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 5370–5381.

[22] R. Sharma, et al., Empathy detection from text, audiovisual, audio or physiological signals: A systematic review, arXiv preprint arXiv:2311.00721 (2023).

[23] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[24] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 9–14.

[25] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 7238–7248.

[26] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the 14th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2020, pp. 1425–1447.

[27] B. Kennedy, X. Jin, A. Mostafazadeh Davani, M. Dehghani, X. Ren, Contextualizing hate speech classifiers with post-hoc explanation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5435–5442.

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

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

[30] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, volume 2, Morgan Kaufmann, 1995, pp. 1137–1143.