Social Support Detection in Social Media: A Multilingual Comparison of Classical ML, BiLSTM, and RoBERTa Approaches

Raúl Ruiz García[1]

[1]Universidad Panamericana, Augusto Rodin 498, Insurgentes Mixcoac, Benito Juárez, 03920, Mexico City, Mexico

Abstract

This paper presents a multilingual system for detecting social support in YouTube comments. The task is structured as three hierarchical subtasks: binary support detection (Task 1), target identification—individual vs. group (Task 2), and multiclass group classification across six communities (Black Community, LGBTQ, Nation, Other, Religion, and Women) (Task 3). We implement and compare three families of models of increasing complexity: classical Machine Learning with TF-IDF features (Logistic Regression and Linear SVM), a Bidirectional LSTM with an attention mechanism trained on fastText embeddings, and fine-tuned Transformer models (twitter-roberta-base for English and BETO for Spanish). We further explore a hybrid strategy that combines RoBERTa for Task 1 with classical ML for Tasks 2 and 3. Our best results on Task 1 are macro-F1 = 0.8423 for Spanish and macro-F1 = 0.8252 for English, both achieved with Transformer-based models. On Task 2, the hybrid strategy yields macro-F1 = 0.5287 (Spanish) and 0.4464 (English). On Task 3, the best results are macro-F1 = 0.5558 (Spanish, BETO) and 0.3905 (English, RoBERTa). Results show that RoBERTa consistently outperforms simpler models on Task 1, while classical ML provides more stable predictions for Tasks 2 and 3 due to the limited size of the hierarchical subsets.

Keywords

social support detection, text classification, transformers, BiLSTM, multilingual NLP, macro-F1

1. Introduction

The analysis of online discourse has been largely dominated by sentiment analysis [1], a paradigm that categorizes text into broad emotional polarities such as positive, negative, or neutral [2]. While useful for capturing general affective tone, this approach falls short of detecting complex prosocial phenomena such as social support, solidarity, or encouragement between users. Social support, expressions of care, admiration, help, or solidarity, is a distinct communicative phenomenon that requires its own detection framework, one that goes beyond polarity labels toward understanding the intent and target of supportive language.

A hierarchical annotation scheme has been proposed for this task, applied to YouTube comments in both English and Spanish. The dataset was introduced by Ahani et al. [3], extended to Spanish by Tash et al. [4], and further explored using advanced ML techniques in Kolesnikova et al. [5] and linguistic analysis in Shahiki-Tash et al. [6]. This work is submitted as part of the SSD-2026 shared task [7] at IberLEF 2026 [8]. To address the three hierarchical subtasks, we implement and compare three model families of increasing representational capacity: (1) **Classical ML** (TF-IDF + Logistic Regression / SVM), chosen as an interpretable and regularization-friendly baseline, especially suited for small training subsets; (2) **BiLSTM with Attention** (Bidirectional Long Short-Term Memory network) trained on fastText embeddings [9], which captures sequential and morphological context; and (3) **Fine-tuned Transformers**, built on the BERT [10] bidirectional pretraining paradigm (RoBERTa [11] for English, BETO [12] for Spanish), which leverage large-scale pretraining on domain-relevant corpora. We additionally propose a hybrid strategy that assigns each model family to the subtask where it performs best.

This work is guided by the following research questions:

IberLEF 2026, September 2026, León, Spain

ruizgar.raul@gmail.com (R. R. García)

- ***RQ1: Which combination of model architecture, language-specific preprocessing, and training strategy best addresses the challenges of hierarchical social support detection in multilingual social media?***
- ***RQ2: To what extent does training data scarcity in hierarchical subtasks undermine the performance of high-capacity models such as Transformers, and under which conditions do simpler models become preferable?***
- ***RQ3: Does the linguistic richness of raw Spanish text provide a measurable advantage over pre-processed English text for fine-grained social support classification tasks?***

Three subtasks are defined: (1) binary detection of whether a comment expresses support, (2) identification of whether that support targets an individual or a group, and (3) multiclass classification of the specific community being supported (Black Community, LGBTQ, Nation, Other, Religion, Women). These subtasks are hierarchically dependent: Task 2 applies only to comments labeled as supportive in Task 1, and Task 3 applies only to group-directed comments from Task 2.

The remainder of this paper is organized as follows. Section 2 reviews related work on sentiment analysis, social support detection, and pre-trained language models. Section 3 describes the task definition and dataset. Section 4 details the preprocessing pipeline. Section 5 presents the methodology, including model architectures and the evaluation metric. Section 6 reports experimental results. Section 7 discusses the findings. Sections 8, 9, and 10 present conclusions, future work, and limitations respectively.

2. Related Work

Social support detection. Ahani et al. [3] introduced the hierarchical annotation framework for social support detection in YouTube comments in English, defining the three subtasks addressed in this work. Using a combination of SVM, BERT, and RoBERTa-based classifiers, they achieved macro-F1 scores above 0.80 on Task 1, establishing the first strong baselines for binary support detection. Their work demonstrated that Transformer models outperform classical approaches on the binary task, a finding we confirm and extend to the bilingual setting. Tash et al. [4] extended this corpus to Spanish, reporting that cross-lingual transfer poses non-trivial challenges due to morphological richness, emojis, and code-switching. Their results showed that monolingual Spanish models outperform zero-shot multilingual transfer for Tasks 2 and 3, motivating our decision to train language-specific models rather than a single multilingual system. Kolesnikova et al. [5] explored a broader set of ML and DL techniques on the same corpus, including gradient boosting, SVM with character n-grams, and fine-tuned Transformers, reporting that Transformer-based models consistently achieved the highest macro-F1 on Task 1, while simpler models remained competitive on Tasks 2 and 3 due to data scarcity. This directly motivates our hybrid strategy.

Sentiment analysis and text classification. The dominance of sentiment analysis in online discourse processing has been extensively documented [1]. Early work established TF-IDF (Term Frequency–Inverse Document Frequency) with SVM as a strong baseline for opinion classification, achieving over 80% accuracy on benchmark datasets [2]. The introduction of word2vec embeddings [13] and later fastText [9], which represents words as character n-gram sums and achieves near state-of-the-art results with significantly faster training, enabled neural architectures to capture richer semantic representations.

Deep learning for text classification. Zhou et al. [14] proposed Attention-Based BiLSTM networks for relation classification, demonstrating that bidirectional context combined with learned attention weights outperforms unidirectional LSTMs by a significant margin on the SemEval relation classification benchmark. This architecture forms the basis of our BiLSTM component. The introduction of the Transformer architecture [15], based entirely on self-attention mechanisms, marked a paradigm shift in NLP by eliminating the sequential bottleneck of recurrent models.

Pre-trained Transformer models. Devlin et al. [10] introduced BERT (Bidirectional Encoder Representations from Transformers), demonstrating that deep bidirectional pretraining followed by task-specific fine-tuning achieves state-of-the-art results across 11 NLP benchmarks. Liu et al. [11]

Table 1
Dataset statistics by language and split.

Dataset	Language	Rows	Split	Columns
trainenglish.csv	English	7,998	Train	id, text, task1–3
trainspanish.csv	Spanish	2,600	Train	id, comment, task1–3
test_phase_english.csv	English	2,000	Test	id, text
test_phase_spanish.csv	Spanish	651	Test	id, comment

showed that longer pretraining with larger batches and more data, producing RoBERTa, consistently improves over BERT, a finding directly relevant to our English model choice. For social media text specifically, Barbieri et al. [16] introduced TweetEval and the `twitter-roberta-base` model, showing that domain-adapted pretraining on Twitter data substantially outperforms generic BERT on seven tweet classification tasks—the model we adopt for English. For Spanish, Cañete et al. [12] introduced BETO, demonstrating competitive or superior performance to multilingual BERT on all Spanish NLP benchmarks evaluated. Nguyen et al. [17] further confirmed that domain-specific pretraining on tweets yields significant gains for social media NLP tasks. For low-resource cross-lingual scenarios, Conneau et al. [18] showed that XLM-RoBERTa, trained on 100 languages, achieves strong results even for languages with limited training data, making it a promising direction for future work on this task.

Prosocial language detection. Hernández Farias et al. [19] studied opinion detection toward refugees in social media, finding that detecting positive and supportive discourse is substantially harder than detecting offensive speech due to greater pragmatic variability and the subtlety of supportive expressions. This finding is consistent with the lower macro-F1 scores observed in our Tasks 2 and 3 compared to Task 1.

The following section formally defines the three subtasks and describes the annotated corpus used in this work.

3. Task and Data Description

3.1. Task Definition

The task defines three hierarchically structured subtasks, evaluated independently using macro-F1 as the official metric:

- **Task 1 – Binary support detection:** Given a social media comment, classify it as *Supportive* (1) or *Non-Supportive* (0). A comment is supportive if it expresses encouragement, care, admiration, help, or solidarity. Comments that promote explicit violence or harm are labeled as Non-Supportive regardless of their apparent tone.
- **Task 2 – Target identification:** Given a supportive comment, classify its target as *Individual* (1) or *Group* (0).
- **Task 3 – Group classification:** Given a group-directed supportive comment, classify the community being supported into one of six categories: Black Community (0), LGBTQ (1), Nation (2), Other (3), Religion (4), Women (5).

All subtasks are evaluated using macro-F1, which assigns equal weight to all classes regardless of frequency, a critical choice given the severe class imbalance present in the data.

3.2. Dataset

The corpus, originally introduced by Ahani et al. [3] and extended to Spanish by Tash et al. [4], consists of YouTube comments annotated in a hierarchical and multilevel fashion. Table 1 summarizes the dataset splits used in this work.

Table 2
Label mapping to numeric encoding.

Subtask	Original label	Numeric value
Task 1	Support / Supportive	1
Task 1	Non-Support / Non-Supportive	0
Task 2	Individual	1
Task 2	Group	0
Task 3	Black Community	0
Task 3	LGBTQ	1
Task 3	Nation / Nacion	2
Task 3	Other / Otro	3
Task 3	Religion	4
Task 3	Women / Mujeres	5
Task 3	No (not applicable)	-1

3.3. Class Distribution and Imbalance

Exploratory data analysis revealed a consistent class imbalance across both languages for Task 1: supportive comments represent only 22.4% of the English training set (1,795 of 7,998) and 22.7% of the Spanish training set (590 of 2,600). This 77/23 split requires explicit class-weighting strategies in all models, as reported in Kolesnikova et al. [5].

The hierarchical structure further reduces the effective training sizes for downstream subtasks. Approximately 1,795 English and 590 Spanish examples are available for Task 2 training, while Task 3 reduces these to roughly 400 and 130 examples respectively, the primary modeling challenge of this work.

For Task 3, class distribution is highly skewed: *Nation* and *Other* account for the majority of examples in both languages, while *Women*, *Religion*, and *Black Community* are severely underrepresented, particularly in Spanish.

Section 4 describes the preprocessing strategies applied to each language before model training.

4. Preprocessing

4.1. Label Mapping

Original string labels exhibited minor variations between languages (e.g., `Supportive` in English vs. `Support` in Spanish). All labels were unified to the numeric encoding required by the evaluation system, as shown in Table 2. Values indicating non-applicability in the hierarchy were encoded as `-1` in submissions.

4.2. Language-Specific Cleaning

An important asymmetry between the two corpora required differentiated preprocessing strategies:

- *English (pre-lemmatized):* The corpus was delivered already tokenized, lowercased, and stripped of stopwords and punctuation, with tokens in their base (lemma) form. Minimal additional cleaning was applied (URL and mention removal) to avoid altering the original preprocessing. A representative example from the training set illustrates this format:

  ```
  sad palestine always support pray indonesia two hearts
  ```

  Note the complete absence of articles, verb inflections, and punctuation—tokens appear in their most reduced form, which limits the sequential and morphological signal available to neural models.

Table 3
Text characteristics by language.

Characteristic	English	Spanish
Mean length (words)	18.7	25.4
Median length (words)	13.0	13.0
Max length (words)	157	744
Std deviation	16.3	42.1
Comments with emoji	0.0%	21.3%
Comments with URL	0.6%	0.1%
Preprocessing state	Lemmatized	Raw

- *Spanish (raw text):* The corpus contains fully natural text with uppercase letters, punctuation, emojis, accented characters, and full morphological variation. The following cleaning pipeline was applied: conversion to lowercase; removal of URLs, @mentions, and #hashtags; emoji removal via ASCII encoding; punctuation stripping while preserving accented characters; whitespace normalization. A representative raw example before cleaning:

  ```
  yo soy mujer pero me gusta mas que me traten como hombre (trans)
  y respeto cualquier otra opinion y orientacion sexual por igual.
  ```

  This example retains natural sentence structure, punctuation, and morphological variation that the English corpus lacks, providing richer signal for models sensitive to contextual and sequential features.

4.3. Text Characteristics

Table 3 summarizes the key quantitative differences between the two corpora after preprocessing. The asymmetry in preprocessing state directly conditioned both the cleaning strategies and the relative performance of each model family across languages, as discussed in Section 7.

Section 5 describes the model architectures and evaluation metric used in this work.

5. Methodology

5.1. Evaluation Metric: Macro-F1

All subtasks are evaluated using the macro-averaged F1 score (macro-F1), the unweighted average of the per-class F1 scores (harmonic mean of precision and recall), which assigns equal weight to all classes regardless of their frequency. This makes it the appropriate metric for imbalanced settings such as this task: a classifier that always predicts *Non-Supportive* would achieve accuracy of 0.77 but macro-F1 of only $\sim$0.44, correctly penalizing the failure to detect the minority class.

5.2. Models

We implement three model families, yielding 18 models in total (2 languages $\times$ 3 subtasks $\times$ 3 families). All models follow the hierarchical structure of the task: separate models are trained for each subtask on the appropriate subset (full training set for Task 1, supportive-only subset for Task 2, group-only subset for Task 3).

5.2.1. Classical ML: TF-IDF + Logistic Regression / SVM

The classical ML pipeline serves as a baseline, following the methodology explored in Kolesnikova et al. [5]. We implemented a `scikit-learn` [20] pipeline with TF-IDF bigrams (`ngram_range=(1,2)`), `sublinear_tf=True`, `max_features=50,000`, and `min_df=2`. We compared Logistic Regression

Table 4
BiLSTM + Attention architecture configuration.

Component	Configuration
Embedding	fastText 300d, fine-tunable
BiLSTM	hidden=128, layers=2, dropout=0.3
Attention	Linear(256,1) + Softmax over sequence
Classifier	Linear(256,64) $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear(64,n)
Optimizer	Adam, lr=10^{-3}, weight_decay=10^{-5}
Scheduler	ReduceLROnPlateau (patience=2, factor=0.5)
Loss	CrossEntropy with inverse-frequency class weights
Epochs	15 with early stopping on val macro-F1

Table 5
Fine-tuning hyperparameters for Transformer models.

Parameter	Task 1	Tasks 2/3
Epochs	5	3
Learning rate	2e-5	1e-5
Batch size	16	32
Warmup steps	10%	10%
Weight decay	0.01	0.01
Dropout	0.3	0.3
Scheduler	Linear warmup + decay	

(`lbfgs`, $C = 1.0$) and Linear SVM (`LinearSVC`, $C = 1.0$), both with `class_weight='balanced'`. The best classifier by macro-F1 in 5-fold stratified cross-validation was selected for test predictions.

5.2.2. Deep Learning: BiLSTM with Attention

We implemented a Bidirectional LSTM with an attention mechanism, trained with PyTorch [21] on GPU (Google Colab T4). We used pre-trained fastText vectors (300 dimensions) [9]: `cc.en.300.vec` for English and `cc.es.300.vec` for Spanish. The architecture consists of a 2-layer bidirectional LSTM [14] with `hidden_dim=128`, `dropout=0.3`, followed by an attention layer and a two-layer classifier head. Table 4 summarizes the full configuration.

5.2.3. Transformers: Fine-tuned RoBERTa / BETO

We selected domain-adapted pre-trained Transformer models [11] for each language:

- **English:** `cardiffnlp/twitter-roberta-base-sentiment` [16], pre-trained on Twitter data.
- **Spanish:** `dccuchile/bert-base-spanish-wwm-cased` (BETO) [12], trained with Whole Word Masking.

A dropout layer ($p = 0.3$) followed by a linear layer maps the `[CLS]` token to target classes. We applied reduced learning rate and fewer epochs for Tasks 2 and 3 after detecting severe overfitting (val_F1$= 1.000$ from epoch 1). All hyperparameters are shown in Table 5.

5.2.4. Hybrid Strategy

Based on subtask-level score analysis, we combine RoBERTa/BETO for Task 1 with classical ML for Tasks 2 and 3, maximizing the strengths of each family as motivated by Kolesnikova et al. [5].

Section 6 presents the experimental results obtained for each model and language.

Table 6
Macro-F1 per subtask, language, and system. Best results per column in bold. Dash (—) indicates score not individually recorded.

Language	System	Task 1	Task 2	Task 3
English	Classical ML	∼0.80	—	—
	BiLSTM + Attention	∼0.80	—	—
	RoBERTa (pure)	0.8252	0.2544	0.3905
	RoBERTa + ML (hybrid)	**0.8252**	**0.4464**	0.3054
Spanish	Classical ML	∼0.79	—	—
	BiLSTM + Attention	∼0.77	—	—
	BETO (pure)	**0.8423**	0.4988	**0.5558**
	BETO + ML (hybrid)	**0.8423**	**0.5287**	0.5526

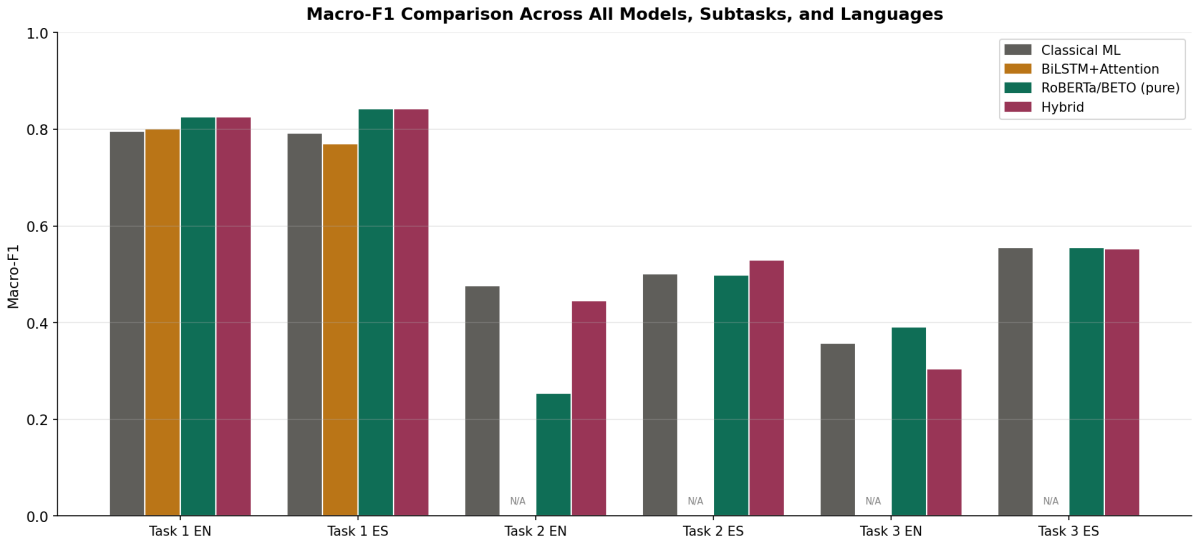

Figure 1: Macro-F1 comparison across all model families, subtasks, and languages. N/A indicates that BiLSTM scores were not recorded individually for Tasks 2 and 3.

6. Results

6.1. Results by Subtask

Table 6 presents the macro-F1 scores per subtask for each system. For classical ML and BiLSTM models, only the global score was recorded during the submission phase; per-subtask scores were tracked from the hybrid submissions onward.[1]

Figure 1 provides a visual overview of macro-F1 scores across all models, subtasks, and languages, with the Task 1 and Tasks 2/3 breakdowns shown together in Figure 2.

6.2. Complete Metrics and Submission Strategy

Table 7 presents the complete evaluation metrics for the pure Transformer and hybrid systems, the two configurations with full subtask coverage. On Task 1, Transformers are the best system for both languages (RoBERTa 0.8252 EN, BETO 0.8423 ES). On Task 2, the hybrid strategy wins for both languages (RoBERTa+ML 0.4464 EN, BETO+ML 0.5287 ES). On Task 3, the pure Transformer wins for both languages (RoBERTa 0.3905 EN, BETO 0.5558 ES); in Spanish the hybrid's classical-ML component trails it by less than one point, but the pure model still edges it out. Across the 11 submissions made,

[1]Global scores for submissions 1–6 correspond to the aggregate macro-F1 reported by the platform across all submitted subtask columns.

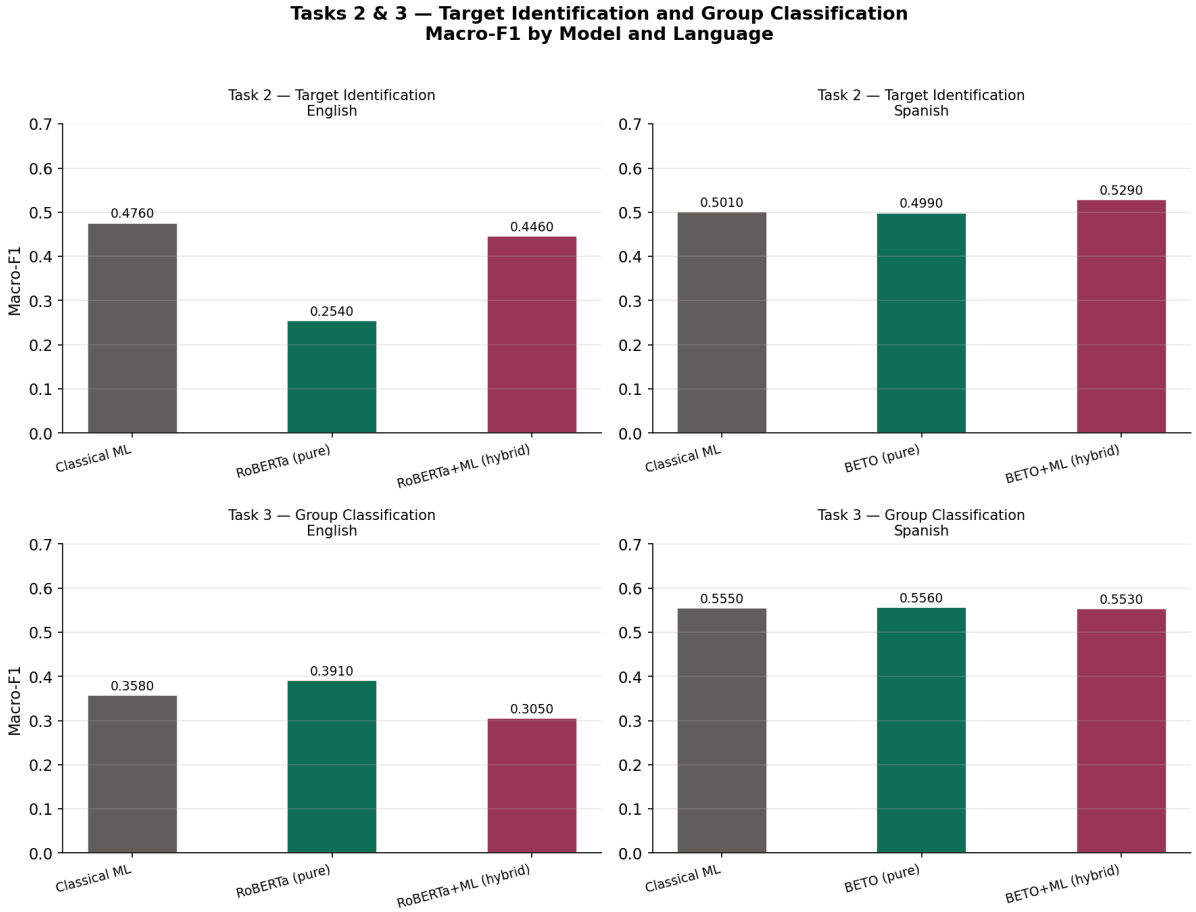

Figure 2: Macro-F1 scores for Task 1 (dashed line marks the 0.80 reference threshold), Task 2 (target identification), and Task 3 (group classification) by model family and language.

the path to these results was incremental: pure single-family submissions established Task 1 baselines around 0.77–0.84 macro-F1; switching Tasks 2–3 to classical ML in the hybrid configuration raised Task 2 macro-F1 sharply over the pure Transformer (e.g., 0.4464 vs. 0.2544 in English); and later attempts to replace the ML component for Task 3 with character n-grams or hand-crafted keywords brought no improvement or actively hurt performance (0.3054 down to 0.1773 in English), confirming that the simple TF-IDF baseline was already the strongest option available for the lower-resource subtasks.

6.3. Error Analysis: Confusion Matrices

Figure 3 shows the confusion matrices for the Classical ML and Transformer models, computed on the official test set labels.

For Task 1, both model families achieve strong performance in both languages, correctly identifying the majority of non-supportive comments. However, both systems show a consistent tendency to misclassify supportive comments as non-supportive (false negatives), reflecting the class imbalance in the test set (approximately 78% non-supportive in English, 77% in Spanish).

For Task 2, Classical ML maintains a more balanced prediction distribution between *Group* and *Individual* targets in both languages. In contrast, RoBERTa collapses entirely to predicting *Group* for all English comments, with zero *Individual* predictions, a direct consequence of the overfitting on the small supportive subset reported in Section 7.

For Task 3, both models concentrate predictions on the dominant classes (*Nation* and *Other* in English; *LGBTQ* in Spanish), failing to recover minority classes such as *Religion*, *Women*, and *Black Community*. This pattern is consistent across both model families and confirms that class imbalance in

Table 7
Complete evaluation metrics for the pure Transformer and hybrid systems, by language and subtask. Best macro-F1 per language–subtask pair in bold.

System	Lang./Task	Acc.	M-F1	W-F1
RoBERTa (pure)	EN / Task 1	0.885	**0.825**	0.882
	EN / Task 2	0.574	0.254	0.628
	EN / Task 3	0.606	**0.391**	0.690
BETO (pure)	ES / Task 1	0.888	**0.842**	0.888
	ES / Task 2	0.650	0.499	0.747
	ES / Task 3	0.545	**0.556**	0.664
Hybrid (RoBERTa+ML)	EN / Task 1	0.885	**0.825**	0.882
	EN / Task 2	0.612	**0.446**	0.732
	EN / Task 3	0.523	0.305	0.617
Hybrid (BETO+ML)	ES / Task 1	0.888	**0.842**	0.888
	ES / Task 2	0.682	**0.529**	0.783
	ES / Task 3	0.545	0.553	0.650

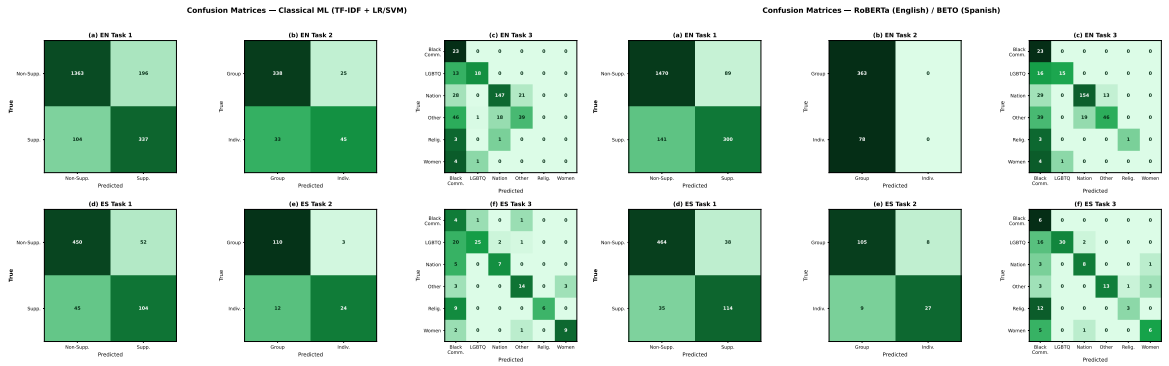

Figure 3: Confusion matrices on the official test set. Left: Classical ML (TF-IDF + LR/SVM). Right: RoBERTa (English) / BETO (Spanish). Within each panel, top row is English and bottom row is Spanish. Color intensity reflects row-normalized frequency; counts show absolute values. Note the complete collapse of Task 2 predictions to a single class for RoBERTa on English, consistent with the overfitting discussed in Section 7.

the hierarchical subsets is the primary bottleneck for Task 3 performance, rather than model capacity per se.

Section 7 interprets these results in relation to the three research questions.

7. Discussion

7.1. RQ1: Best Combination of Architecture, Preprocessing, and Strategy

Addressing RQ1, the results show that no single model family dominates across all subtasks. RoBERTa and BETO consistently outperform both the classical ML and BiLSTM baselines on Task 1 by 3–5 macro-F1 points (Figure 2), confirming that global contextual representations [11] are essential for detecting social support. For Tasks 2 and 3, however, classical ML [20] proves more stable. The optimal combination is the hybrid strategy: Transformer for Task 1, classical ML for Tasks 2 and 3. The confusion matrices (Figure 3) confirm this pattern visually: both model families show strong diagonal dominance on Task 1, while Task 3 predictions concentrate on *Nation* and *Other* in English and *LGBTQ* in Spanish, with near-zero recall for minority classes.

7.2. RQ2: Data Scarcity and Model Capacity

Addressing RQ2, the reduced subset sizes (∼400 examples for Task 3 in English, ∼130 for Spanish) create severe instability during Transformer fine-tuning: RoBERTa and BETO overfit from the first epoch (val_F1 $= 1.000$) and subsequently predict a single class on the test set. This confirms that high-capacity models become disadvantageous when data is scarce. A keyword-based fallback for Task 3 further degraded performance (from 0.3054 to 0.1773 for English), as generic keywords caused inter-class conflicts especially with the *Other* class. This collapse is directly visible in the right panel of Figure 3, where the English Task 2 confusion matrix shows all test instances predicted as *Group*, yielding zero recall for *Individual*. Classical ML avoids this failure mode (left panel), maintaining non-trivial predictions for both classes, which explains its superior Task 2 macro-F1 despite lower Task 1 performance.

7.3. RQ3: Linguistic Richness of Spanish vs. Pre-processed English

Addressing RQ3, despite having significantly less training data (2,600 vs. 7,998 examples), Spanish consistently achieves higher scores than English on Tasks 2 and 3 (Figure 2). BETO pure achieves macro-F1 = 0.5558 on Spanish Task 3 vs. 0.3905 for RoBERTa on English Task 3 (Table 7). This suggests that raw Spanish text—with its morphological richness, emojis, and pragmatic cues—is more informative for community identification, consistent with the findings of Tash et al. [4]. The real examples in Section 4.2 further illustrate this contrast: the English corpus loses verb inflections, articles, and sentence structure that the Spanish corpus preserves. The Task 3 confusion matrices further support this finding: the Spanish model achieves non-zero recall across more community categories than its English counterpart, despite having three times fewer training examples. This suggests that morphological richness in raw Spanish text compensates partially for data scarcity in fine-grained community classification.

Section 8 summarizes the main conclusions of this work.

8. Conclusions

This paper presented a multilingual system for detecting social support in YouTube comments [3, 4] across three hierarchical subtasks. Addressing RQ1, the optimal approach is a hybrid strategy combining Transformer models for Task 1 with classical ML for Tasks 2 and 3. Addressing RQ2, Transformer fine-tuning is unreliable with fewer than ∼400 training examples, where classical ML is preferable. Addressing RQ3, raw Spanish text provides measurably more signal than pre-processed English for fine-grained community classification despite fewer training examples.

The best results on Task 1 are macro-F1 = 0.8252 (English, RoBERTa) and macro-F1 = 0.8423 (Spanish, BETO). On Task 2, the hybrid strategy achieves macro-F1 = 0.4464 (English) and 0.5287 (Spanish). On Task 3, the best results are macro-F1 = 0.3905 (English, RoBERTa) and 0.5558 (Spanish, BETO).

9. Future Work

Several directions could substantially improve performance. Multilingual training with XLM-RoBERTa [18] would effectively double the Task 3 training set. Data augmentation via automatic translation could increase examples for minority classes (*Women*, *Religion*, *Black Community*). Ensemble methods combining all three model families may improve robustness on Tasks 2 and 3. Few-shot prompting with large language models represents a promising alternative for Task 3 where fine-tuning is data-insufficient. Finally, soft hierarchical propagation using predicted class probabilities instead of hard labels could reduce compounding cascade errors between subtasks.

10. Limitations

Dataset limitations. The training subsets for Tasks 2 and 3 are extremely small (∼130–400 examples), making reliable model training and evaluation difficult. The class distribution in Task 3 is highly skewed, with *Women*, *Religion*, and *Black Community* severely underrepresented, particularly in Spanish, inflating variance in per-class F1 estimates. Additionally, the asymmetric preprocessing between languages (lemmatized English vs. raw Spanish) introduces a confound when comparing cross-lingual results.

Model limitations. The Transformer models (∼110–125M parameters) are susceptible to overfitting on small subsets even with reduced learning rates and early stopping. The BiLSTM is constrained by the fixed fastText vocabulary and the absence of subword tokenization, limiting its handling of morphological variants in Spanish. The classical ML pipeline relies on surface-level n-gram features that may miss pragmatic nuances essential for Tasks 2 and 3. Finally, the hierarchical pipeline propagates prediction errors from Task 1 through Tasks 2 and 3, amplifying mistakes at each level.

Acknowledgments

The author thanks the organizers of the SSD-2026 Shared Task on Social Support Detection [7] at IberLEF 2026 [8] for providing the dataset and evaluation infrastructure. Special thanks to the authors of Ahani et al. [3], Tash et al. [4], Kolesnikova et al. [5], and Shahiki-Tash et al. [6] for making the annotated corpus and related analyses publicly available.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: writing assistance, editing, and code generation support. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012. doi:10.2200/S00416ED1V01Y201204HLT016.

[2] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86. doi:10.3115/1118693.1118704.

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

[4] 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.

[5] 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). doi:10.1016/j.heliyon.2025.e43437.

[6] 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). Preprint. https://doi.org/10.21203/rs.3.rs-9557287/v1.

[7] M. Shahiki Tash, L. Ramos, A. Zahra, 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.

[8] 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.

[9] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 427–431. doi:10.18653/v1/E17-2068.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] 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). URL: https://arxiv.org/abs/1907.11692.

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Proceedings of the PML4DC Workshop at ICLR 2020, 2020.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, 2013.

[14] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 207–212. doi:10.18653/v1/P16-2034.

[15] 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, 2017.

[16] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1644–1650. doi:10.18653/v1/2020.findings-emnlp.148.

[17] D. Q. Nguyen, T. Vu, A. Tuan 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, 2020, pp. 9–14. doi:10.18653/v1/2020.emnlp-demos.2.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[19] D. I. Hernández Farias, V. Patti, P. Rosso, Detecting positive and negative opinions toward refugees in social media, in: Proceedings of the Workshop on Abusive Language Online, 2021, pp. 1–10.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019.