Social Support Detection in YouTube Comments: A Hierarchical Transformer Approach with Class-Balanced Focal Loss and LLM-Based Augmentation

Daniel Alejandro Osornio López[1,*], Josué Jyaru Hernandez Carmona[1] and Moein Shahiki Tash[1]

[1]*Universidad Panamericana, Ciudad de México, México*

Abstract

Automatically detecting social support in online discourse presents unique challenges: the task requires modeling communicative intent rather than surface sentiment, and real-world datasets exhibit severe class imbalance across fine-grained categories. We study this problem on YouTube comments in Spanish and English, structured as three hierarchical subtasks: (1) supportive vs. non-supportive detection, (2) individual vs. group-targeted support, and (3) identification of the specific group target among six categories. We investigate the relative contribution of psycholinguistic features, high-dimensional API embeddings, and multilingual transformer representations, and propose a multi-stage pipeline combining Class-Balanced Focal Loss ($\beta = 0.9$, $\gamma = 2$) for transformer fine-tuning, LLM-based minority class augmentation, and probability-level late fusion with grid-searched weights. Our experiments show that 108-dimensional psycholinguistic features consistently outperform embeddings of up to 4096 dimensions, that combining transformer fine-tuning with CB-Focal loss improves Task 3 from 0.52 (best traditional ML) to 0.86–0.92 (transformer with CB-Focal), and that for Task 1 a weighted ensemble of two domain-specific transformers outperforms any single model. On the held-out test set, we obtained macro F1 of **0.8291** (EN) / **0.8707** (ES) for Task 1, **0.8474** (EN) / **0.8978** (ES) for Task 2, and **0.8605** (EN) / **0.9211** (ES) for Task 3 (per-task averages: 0.8499 / 0.8726 / 0.8908).

Keywords

social support detection, hierarchical classification, class imbalance, multilingual NLP, late fusion ensemble

1. Introduction

Social support, defined as the provision of emotional, informational, or instrumental assistance through social relationships, is a fundamental human need with well-documented benefits for mental health, well-being, and resilience [1]. In the digital age, social media platforms have become primary venues where individuals seek and receive support, particularly during crises, social movements, and personal challenges. YouTube, as one of the largest user-generated content platforms, hosts millions of comments where users express solidarity, offer advice, or simply acknowledge others' struggles. Detecting such supportive language is a significant Natural Language Processing (NLP) challenge with practical applications in community moderation, mental health triage, and social science research.

From an NLP perspective, social support detection presents unique challenges distinct from related tasks such as sentiment analysis or emotion detection. While sentiment analysis classifies text polarity and emotion detection identifies affective states, support detection requires understanding the functional role of a message within a social interaction. A comment like "You are so brave, keep fighting!" is clearly supportive, while "This video is trash" is clearly not—yet both are straightforward for sentiment analysis. The challenge lies in the vast middle ground: comments that are warm but impersonal, critical but caring, or ambiguous in intent. This pragmatic dimension necessitates models that capture interactional context and communicative intent beyond surface-level affect.

The dataset consists of 10,600 YouTube comments [2, 3, 4, 5] from the Social Support Detection shared task at IberLEF 2026 [6, 7], labeled across three hierarchical subtasks of increasing specificity:

IberLEF 2026, León, Spain

*Corresponding author.

0244685@up.edu.mx (D. A. O. López); 0246759@up.edu.mx (J. J. H. Carmona); mshahiki@up.edu.mx (M. S. Tash)

1. **Support Detection**: Binary classification of comments as supportive or non-supportive.
2. **Target Identification**: For supportive comments, distinguishing between individual-targeted support (directed at a specific person) and group-targeted support (directed at a demographic or identity-based community).
3. **Group Target Classification**: For group-targeted support, identifying the specific community being supported among six categories: Nation, LGBTQ, Black Community, Women, Religion, and Other.

The dataset presents significant challenges for classification. The comments are drawn from politically and socially charged video contexts, featuring code-mixed Spanish and English text with informal register, slang, emoji usage, and platform-specific conventions. Most critically, the distribution of classes is highly imbalanced, particularly at the deepest level of the hierarchy: some group target categories contain fewer than 20 examples in the training data, while others exceed 700.

In this paper, we present a multi-stage pipeline addressing these challenges through psycholinguistic features, domain-specific transformer fine-tuning with Class-Balanced Focal Loss, LLM-based minority class augmentation, and late fusion ensemble methods. We compare domain-adapted models pre-trained on social media text—BERTweet [8] for English and RoBERTuito [9] for Spanish—against general-purpose multilingual transformers across all three subtasks. On the held-out test set, we obtained average macro F1 of **0.8499**, **0.8726**, and **0.8908** for Tasks 1, 2, and 3 respectively (averaged across English and Spanish).

This work is guided by the following research questions:

- **RQ1:** Do compact psycholinguistic features (108 dimensions) capture social support signals more effectively than high-dimensional general-purpose embeddings (768–4096 dimensions) for this task?
- **RQ2:** Does Class-Balanced Focal Loss provide meaningful gains over standard cross-entropy for fine-grained group classification under extreme class imbalance, and how sensitive is performance to the choice of the β hyperparameter?
- **RQ3:** Does domain-specific social media pretraining (BERTweet, RoBERTuito) outperform general multilingual capacity (mBERT, XLM-RoBERTa), and does this advantage vary across task granularity?
- **RQ4:** Does late fusion of transformer and psycholinguistic-based traditional models improve over either component alone, and which tasks benefit most?

The remainder of this paper is organized as follows. Section 2 reviews related work on social support detection, hierarchical classification, class imbalance, and multilingual NLP. Section 3 describes our pipeline in detail. Section 4 presents experimental results. Section 5 discusses findings and limitations. Section 6 concludes.

2. Related Work

2.1. Social Support as an NLP Task

Social support, defined by House [1] as the provision of emotional, informational, instrumental, or appraisal assistance through social relationships, has been studied extensively in psychology. Cutrona and Russell [10] established a fine-grained taxonomy of support types—emotional, informational, tangible aid, and esteem support—and found through diary studies that the optimal match between support type and stressor nature predicts recipient satisfaction. This theoretical grounding motivates our hierarchical labeling scheme: rather than treating support as monolithic, we distinguish whether it targets an individual or a group and, if the latter, which community.

NLP research on social support has largely focused on health-related forums and mental health communities. Yang and Kraut [11] modeled social roles in online health communities (Reddit r/depression,

r/cancer), identifying four behavioral archetypes: seekers, providers, welcomers, and storytellers. Using text classification and linguistic feature analysis, they found that role enactment is predictable from linguistic cues and that the same user frequently transitions between roles over time. This work established that social support detection requires modeling pragmatic and interactional dimensions beyond surface sentiment, a challenge our approach addresses through psycholinguistic features and intent-sensitive transformer fine-tuning.

Our dataset and task definition stem directly from a line of work by Ahani et al. [3], Shahiki Tash et al. [2], Kolesnikova et al. [4]. Ahani et al. [3] introduced the three-level hierarchical annotation scheme on 10,000 YouTube comments in English and Spanish, achieving best macro F1 scores between 0.72 and 0.82 across subtasks using combinations of n-gram features with psycholinguistic, emotional, and sentiment signals. Their finding that group-oriented support predominates in politically charged comment sections—and that fine-grained group classification is where models fail most—directly motivated our application of Class-Balanced Focal Loss in Stage 3. Building on this, Shahiki Tash et al. [2] constructed a Spanish-only annotated corpus of 3,189 YouTube comments spanning the same three subtasks, finding that balanced training data significantly improves Tasks 2 and 3 (group target identification), while GPT-4o achieved the strongest results on Task 1 binary support classification. Kolesnikova et al. [4] extended this with ensemble methods and advanced resampling strategies for social support detection on social media, providing further evidence that naive single-model approaches are insufficient under the extreme class imbalance present in Tasks 2 and 3.

2.2. Hierarchical Text Classification

Hierarchical classification, where class labels form a taxonomy and predictions at one level constrain predictions at deeper levels, has been surveyed by Silla and Freitas [12] across biology, text categorization, and music information retrieval. Their survey distinguishes two broad strategies: *local classifiers*, which train one model per node in the taxonomy and propagate predictions top-down, and *global classifiers*, which optimize a single model across all levels jointly. Local classifiers offer the advantage of task-specific model selection but risk error propagation, while global classifiers enable multi-task regularization at the cost of architectural complexity. Silla and Freitas [12] found that the optimal strategy is highly domain-dependent: for sparse, imbalanced taxonomies—precisely our setting—local classifiers per level tend to outperform global models because they allow independent handling of each level's class distribution.

Our pipeline adopts a local classifier design, training independent models per subtask on the hierarchically filtered subsets provided by the ground-truth labels (Task 2 trains on supportive-only rows; Task 3 trains on group-only rows). At inference time, however, we predict all three tasks for every comment unconditionally, without cascading Task 1 predictions into Task 2 or Task 3. This decoupling is motivated empirically: 29 comments in the training data carry group-target labels despite being labeled non-supportive in Task 1, indicating that the annotation layers are not strictly nested. Predicting independently also avoids error propagation from Task 1 into deeper subtasks. This design additionally allows task-specific model selection without committing to a single backbone for all levels.

2.3. Addressing Class Imbalance in Deep Learning

Class imbalance is a pervasive challenge in fine-grained NLP, particularly when modeling rare categories. Lin et al. [13] introduced Focal Loss to address the extreme foreground-background imbalance in object detection. Focal Loss modifies cross-entropy by a multiplicative factor $(1 - p_t)^\gamma$ that down-weights easy examples exponentially, allowing the model to focus on hard examples. Although originally designed for computer vision, Focal Loss transfers naturally to severely imbalanced text classification.

Cui et al. [14] extended this with Class-Balanced (CB) Focal Loss, arguing that raw sample counts are a poor proxy for the information content of a class. They propose the *effective number of samples* $E_n = (1-\beta^n)/(1-\beta)$ where $\beta \in [0, 1)$ controls how quickly marginal information diminishes with addi-

tional samples. Each class is reweighted inversely proportional to its effective number, then combined with Focal Loss. CB-Focal consistently outperforms raw Focal Loss, class-frequency reweighting, and oversampling baselines on long-tailed benchmarks. Our Task 3 exhibits comparable imbalance ratios (Nation at n>700 vs. Religion at n=15), motivating our adoption of CB-Focal with $\beta = 0.9$.

2.4. Multilingual and Domain-Specific Transformers

Devlin et al. [15] introduced BERT, which pre-trains deep bidirectional representations through masked language modelling (MLM) and next sentence prediction (NSP). Fine-tuned BERT achieved state-of-the-art results across a wide range of NLP tasks. The multilingual variant (mBERT), trained on 104 languages, enables cross-lingual transfer with a single model—making it a natural baseline for our bilingual Spanish-English task.

Conneau et al. [16] scaled multilingual pretraining with XLM-RoBERTa (XLM-R), training a RoBERTa-style model on 2.5TB of filtered CommonCrawl text across 100 languages. XLM-R-Large substantially outperforms mBERT on cross-lingual benchmarks. Crucially, Conneau et al. demonstrate that larger models trained on larger multilingual corpora do not suffer the "curse of multilinguality" that limits smaller models, making XLM-R-Large our primary backbone for Tasks 2 and 3 where cross-lingual capacity is most valuable.

For Spanish, Cañete et al. [17] released BETO, a BERT-base model pre-trained on the Spanish Wikipedia and a curated Spanish web corpus. BETO outperforms mBERT on a suite of Spanish NLP benchmarks, establishing that dedicated monolingual pretraining provides advantages over multilingual models for in-language tasks. Pires et al. [18] provide a systematic analysis of mBERT's multilingual representations, finding that the model does learn genuinely cross-lingual features, but that its cross-lingual transfer is strongest between typologically similar languages and degrades for more distant language pairs, reinforcing the value of monolingual or domain-specific alternatives when available.

Beyond language coverage, domain pretraining matters substantially for informal text. Nguyen et al. [8] released BERTweet, the first BERT-base model pre-trained on a large English tweet corpus (873 million tweets). On downstream Twitter NLP tasks, BERTweet consistently outperforms both BERT-base and RoBERTa-base, attributed to domain-specific vocabulary and linguistic patterns. This is particularly relevant for our dataset, where YouTube comments share the informal register, abbreviations, and emoji usage characteristic of social media text.

Pérez et al. [9] developed RoBERTuito, a RoBERTa-based model pre-trained on over 500 million Spanish tweets, and evaluated it on a benchmark of Spanish social media tasks. RoBERTuito outperformed BETO and all other available Spanish language models on these benchmarks, demonstrating that domain specialisation for social media content is at least as important as language coverage for tasks involving user-generated text. This finding directly motivates our use of RoBERTuito for Spanish subtasks in preference to BETO.

2.5. Psycholinguistic Features and Neural Models

Fast et al. [19] introduced Empath, a tool for generating psycholinguistic features from text across 194 thematic categories (ranging from affective categories such as *anger* and *joy* to topical categories such as *religion* and *politics*). Unlike LIWC, which relies on a static manually curated lexicon, Empath generates category lexicons by seeding with example words and expanding them using word embeddings. Validated against LIWC on Reddit corpora, Empath achieves high correlation with LIWC category scores while offering greater flexibility and coverage. Its 194 dimensions provide a compact but semantically rich feature space for capturing the topical and emotional content of social support language.

Pérez et al. [20] released pysentimiento, an open-source Python toolkit providing pre-trained models for sentiment analysis, emotion recognition, irony detection, and hate speech across Spanish, English, Italian, and Portuguese. The sentiment and emotion models are fine-tuned transformer classifiers for

social media text, outputting calibrated class probabilities that can be used directly as features, making pysentimiento a well-validated source of affective signal for our 108-dimensional psycholinguistic feature vector.

A recurring finding in the literature is that naive feature concatenation of psycholinguistic signals with transformer representations does not reliably improve performance: the mismatch in scale, dimensionality, and distributional properties between handcrafted feature vectors and contextual embeddings can destabilize training. We design our integration accordingly, fusing psycholinguistic-based traditional classifiers with transformer models at the probability level rather than the feature level, allowing each component to be independently calibrated before combination.

3. Methodology

Our approach follows a staged pipeline architecture, with each stage producing artifacts consumed by subsequent stages. This modularity enables systematic ablation and debugging while ensuring reproducibility.

3.1. Data Preprocessing and Feature Extraction

The dataset comprises 10,598 YouTube comments distributed across English (7,998) and Spanish (2,600) with significant code-mixing. Our preprocessing pipeline addresses platform-specific noise while preserving affective signals: URLs and @mentions are removed, excessive whitespace is normalized, but emojis are retained and demojized as tokens (e.g., `:smiling_face:`) since they carry significant emotional content in informal online communication.

We extract four categories of features for our baseline experiments, of which only psycholinguistic features are carried forward to the final ensemble:

1. **TF-IDF**: Word n-grams (1–2) with a maximum of 50,000 features, reduced to 300 dimensions via TruncatedSVD for tree-based models.
2. **Distributed Representations**: Pre-trained Word2Vec embeddings (300-dimensional, CBOW, window=5) averaged per document.
3. **API Embeddings**: Embeddings from Nomic (768-d), Gemini (3072-d), Qwen3 (4096-d), and OpenAI text-embedding-3-large (3072-d) obtained via API calls with SHA256-based caching.
4. **Psycholinguistic Features**: 108-dimensional vectors combining sentiment (neutral/positive/negative) and seven emotion probabilities (anger, disgust, fear, joy, sadness, surprise, trust) from pysentimiento [20] with 98 psycholinguistic categories from Empath [19].

All six feature representations exhibit substantial class overlap in t-SNE projections, motivating our use of psycholinguistic features for the traditional ML component.

Label encoding is performed per (language, task) pair using scikit-learn's `LabelEncoder`, with encoders serialized to disk for reproducible decoding of predictions.

Table 1 summarises the dataset and class distributions per task and language. The annotation structure is hierarchical: Task 2 is annotated only for supportive comments (2,385 of 10,598) and Task 3 includes all supportive comments (2,385), with group-targeted rows forming a subset (1,935 of 2,385). Task 3 models are trained on all 2,385 supportive rows, where the "No" class (individual-level support, 345 EN + 134 ES = 479 rows) is included as a distinct category. At inference time, all three tasks are predicted independently for every comment, without cascading Task 1 predictions into Task 2 or Task 3.[1] Severe imbalance is evident in Task 3, where Religion (EN=15) has 52× fewer examples than Nation (EN=786).

[1] 29 training examples carry group-target labels despite being labeled Non-Supportive in Task 1, indicating that the annotation layers are not strictly nested. Predicting unconditionally also avoids error propagation from Task 1 into deeper subtasks.

Table 1
Dataset size and class distribution summary by task and language. **Bold** = most frequent class; *italic* = least frequent class per language block.

Task	Filter	Lang	Class	Count	%
Task 1	all data	EN	**Non-Supportive**	**6,203**	**77.6**
			Supportive	*1,795*	*22.4*
		ES	**Non-Support**	**2,010**	**77.3**
			Support	*590*	*22.7*
Task 2	supportive	EN	**Group**	**1,450**	**80.8**
			Individual	*345*	*19.2*
		ES	**Group**	**456**	**77.3**
			Individual	*134*	*22.7*
Task 3	group	EN	**Nation**	**786**	**54.2**
			Other	416	28.7
			LGBTQ	123	8.5
			Black Community	91	6.3
			Women	19	1.3
			Religion	*15*	*1.0*
		ES	**LGBTQ**	**224**	**46.2**
			Other	82	16.9
			Religion	60	12.4
			Women	48	9.9
			Nation	47	9.7
			Black Community	*24*	*4.9*

3.2. Traditional ML Baselines

We evaluate four traditional classifiers—Logistic Regression, LinearSVC, Random Forest, and XGBoost—under 5-fold cross-validation using each feature source independently. This systematic ablation reveals which feature representations carry discriminative signal for social support, before committing to the more expensive transformer fine-tuning stage.

XGBoost uses GPU-accelerated tree construction (`tree_method=hist`, CUDA device) with `scale_pos_weight` adjusted per class for Tasks 1 and 2. All classifiers use `random_state=42`. The best-performing feature-classifier combination per (language, task) is selected as the traditional ML component of the late-fusion ensemble.

3.3. Class-Balanced Focal Loss

For Task 3, where class imbalance is most severe, we employ Class-Balanced Focal Loss (CB-Focal) [14]. Given n_y samples for class y, the effective number is defined as:

$$E(n_y) = \frac{1 - \beta^{n_y}}{1 - \beta} \tag{1}$$

where $\beta \in [0, 1)$ is a hyperparameter. The class-balanced weight is:

$$w_y = \frac{1 - \beta}{1 - \beta^{n_y}} \tag{2}$$

The focal loss component modulates the standard cross-entropy with a focusing parameter $\gamma = 2$:

$$\text{CB-Focal}(p_t) = -w_y(1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

where p_t is the predicted probability for the true class.

A critical implementation detail is the choice of β. The original formulation [14] suggests $\beta = 0.9999$, but we found this produces near-zero weights for the majority class in practice: as $n_y \to \infty$, $\beta^{n_y} \to 0$ and $w_y \to (1 - \beta) \approx 0.0001$, which drives training loss to $\sim 10^{-4}$ and gradients to $\sim 10^{-14}$. We instead use $\beta = 0.9$, which maintains the relative weighting between classes while allowing healthy gradient magnitudes (loss in the 0.01–0.08 range). Under this setting, the 52× sample-count imbalance between the majority class (Nation: 786 samples) and the rarest class (Religion: 15 samples) yields a CB-Focal weight ratio of approximately 1.26× ($w_{\text{Religion}}/w_{\text{Nation}} = 0.126/0.100$), providing enough signal without extreme gradient scaling. CB-Focal is applied to all transformer fine-tuning experiments across all three tasks. Tasks 1 and 2, while less severely imbalanced, also benefit from the focal component's emphasis on hard examples.

3.4. Transformer Fine-Tuning

We evaluate five transformer architectures across two phases of experimentation:

Phase A — Multilingual and monolingual baselines:

Multilingual mBERT (`google-bert/bert-base-multilingual-cased`, 179M parameters) is trained jointly on English and Spanish data, leveraging cross-lingual transfer for tasks where per-language data is limited.

Spanish BERT / BETO (`dccuchile/bert-base-spanish-wwm-cased`, 110M parameters) is trained on Spanish-only data. The whole-word-masking cased variant preserves orthographic cues important for informal text.

Phase B — Stronger and domain-specific backbones:

XLM-RoBERTa-Large (`FacebookAI/xlm-roberta-large`, 560M parameters) is trained jointly on English and Spanish. Its larger capacity is beneficial for the nuanced group-level classification in Tasks 2 and 3, but we observe overfitting on the simpler binary Task 1.

BERTweet (`vinai/bertweet-base`, 135M parameters) is pre-trained on 873 million English tweets. Its domain match to the informal register of YouTube comments makes it a strong English backbone for Task 1.

RoBERTuito (`pysentimiento/robertuito-base-uncased`, 125M parameters) is pre-trained on over 500 million Spanish tweets, providing the same domain advantage for Spanish.

Training configuration: All models use maximum sequence length 128, AdamW optimizer with weight decay 0.01, linear warmup (10% of steps) followed by cosine decay, early stopping with patience 5, and a maximum of 20 epochs. Batch size is 16; XLM-RoBERTa-Large uses batch size 8 with 2-step gradient accumulation for an effective batch of 16. Learning rates are 2e-5 for all base-size models and 1e-5 for XLM-RoBERTa-Large. All experiments use random seed 42. Training runs on an NVIDIA RTX 5090 GPU (32 GB VRAM).

3.5. LLM-Based Data Augmentation

To address severe class imbalance in Task 3, we define minority classes as those with fewer than 100 training samples in the group-targeted subset. The affected classes are: EN Black Community=91, Religion=15, Women=19; ES Black Community=24, Nation=47, Other=82, Women=48, Religion=60. We generate synthetic training examples using three large language models via API: GLM-4 [21] (Synthetic API), Claude Sonnet (OpenRouter), and Gemini 2.0 Flash (OpenRouter). For each minority class we generate 20 same-intent paraphrases per model where available (one class-language combination received only 2 of 3 models due to generation failures), resulting in 460 synthetic samples across 8 underrepresented class-language combinations. Figure 1 shows the resulting class distributions before and after augmentation.

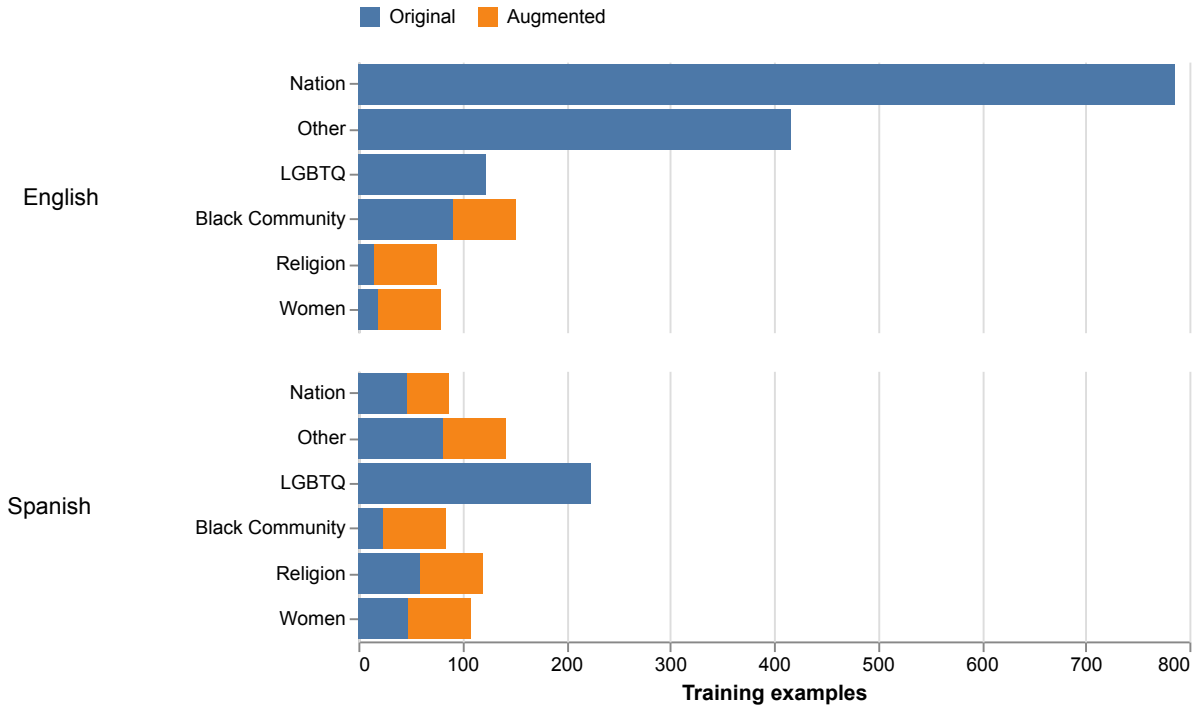

Figure 1: Task 3 class distribution before and after LLM augmentation for English (left) and Spanish (right). Augmented samples from three LLMs (Claude, Gemini, GLM-4) are stacked atop the original counts for each minority class.

Generation prompts include explicit language instructions ("write this comment in English" or "escribe este comentario en español") to ensure code-mixed prompts produce monolingual outputs appropriate to the target class. Quality is verified via TF-IDF cosine similarity to real examples; Spanish generations achieve 0.16–0.18 mean similarity after prompt refinement, comparable to English. Augmented samples are included only in training folds, never in validation. English Task 3 improves by +0.045 macro F1 after augmentation, while Spanish regresses by −0.017, suggesting the Spanish prompt may not capture social support patterns as effectively.

3.6. Ensemble Strategy

Model selection and fusion are guided by macro F1 on a fixed 20% held-out validation split. Our final system uses a two-level strategy: (1) *per-task model selection*, where each subtask independently selects its best-performing model(s) from the validation set, and (2) *within-task probability fusion*, where for tasks with no single dominant model we fuse multiple models' probability outputs via a grid-searched weighted average.

Task 1 — Weighted transformer ensemble. We find that no single model dominates Task 1 across both languages. For English, BERTweet and mBERT achieve near-identical validation performance; their probability-level weighted average (BERTweet 50.4%, mBERT 49.6%) outperforms either alone. For Spanish, RoBERTuito and Spanish BERT are similarly matched (RoBERTuito 50.7%, Spanish BERT 49.3%). Weights are determined by grid search over the held-out validation set.

Tasks 2 & 3 — Single best model. XLM-RoBERTa-Large trained jointly on English and Spanish outperforms all other models on both tasks and both languages, making ensemble fusion unnecessary.

Task 3 label remapping. A critical implementation detail: Task 3 true labels are non-contiguous ($\{0, 1, 2, 4, 5, 6\}$ in the original encoding; class 3 "No" is absent because Task 3 operates only on group-targeted rows). Transformers are trained with contiguous labels $[0, \ldots, 5]$ via an explicit remapping. At inference time, the "No" class column is removed from the 7-class output distribution and probabilities are renormalized before taking the argmax. Without this fix, label misalignment causes Task 3 macro F1 to collapse to 0.37–0.42 despite correct training.

4. Results

4.1. Traditional Baseline Performance

Table 2 summarises the best traditional ML score per feature source for Tasks 1 and 2 (Task 3 omitted: all methods score below 0.52 due to extreme imbalance). The full ablation grids across all feature sources, classifiers, and tasks are provided in the Appendix (Tables 6–8), with complete validation metrics for the strongest traditional setting in Table 9.

Psycholinguistic features demonstrate clear dominance across Tasks 1 and 2, with XGBoost and LinearSVC achieving the best performance. API embeddings underperform psycholinguistic features by a large margin (e.g., 0.7354 vs 0.4882 for EN Task 1 best XGBoost vs best Gemini), indicating that general-purpose embedding spaces introduce noise that overwhelms the carefully curated psycholinguistic signal: 108-dimensional psycholinguistic features achieve 0.73–0.74 macro F1 on Task 1, while 768–4096-dimensional API embeddings cluster below 0.50. Task 3 results remain near-random (0.10–0.32) for all traditional methods, motivating the use of Class-Balanced Focal Loss and transformer fine-tuning.

Table 2
Best traditional ML result per feature source (best model across LR/SVC/RF/XGB). Format: Macro F1 / Weighted F1. Task 3 omitted: all methods score below 0.52 due to extreme imbalance.

	Task 1		Task 2	
Feature Source	**EN**	**ES**	**EN**	**ES**
Psycholinguistic	**0.7354** / 0.8134	**0.7906** / 0.8527	**0.6836** / 0.8093	**0.6577** / 0.7608
TF-IDF	0.5011 / 0.6593	0.4773 / 0.6735	0.4720 / 0.7159	0.4490 / 0.6735
Word2Vec	0.4976 / 0.6738	0.4832 / 0.6742	0.4871 / 0.7183	0.4909 / 0.6653
Nomic	0.4797 / 0.6850	0.4784 / 0.5880	0.4876 / 0.6342	0.5075 / 0.6235
Gemini	0.4882 / 0.5914	0.4789 / 0.5978	0.4990 / 0.6614	0.5293 / 0.6626
Qwen3-8B	0.4797 / 0.5849	0.4766 / 0.5950	0.4939 / 0.6538	0.5110 / 0.6552
OpenAI-Large	0.4886 / 0.5894	0.4864 / 0.5971	0.5093 / 0.6631	0.5112 / 0.6563

4.2. Transformer Performance

Table 3 compares all five transformer architectures evaluated across the two experimental phases. Models trained jointly on EN+ES are evaluated on the joint held-out split (2,120 samples); monolingual models are evaluated on their language-specific split (EN: 1,600 samples; ES: 520 samples).

Table 3
Transformer Fine-Tuning Results (Best Validation Macro F1). All experiments use CB-Focal β=0.9, γ=2. Dashes indicate tasks for which the model was not fine-tuned.

Model	**Training**	**Task 1**	**Task 2**	**Task 3**
Phase A — Multilingual and monolingual baselines				
mBERT	joint EN+ES	0.7649	0.8474	0.7994
Spanish BERT	ES only	**0.8437**	0.7393	0.7814
Phase B — Stronger and domain-specific backbones				
XLM-RoBERTa-Large	joint EN+ES	0.7611	**0.8631**	**0.8195**
BERTweet	EN only	**0.8039**	—	—
RoBERTuito	ES only	**0.8692**	0.8374	0.8455

Results reveal clear task-dependent patterns. For Task 1 (binary support detection), domain-specific social media pre-training dominates: BERTweet (English Twitter, 873M tweets) achieves 0.8039 on English and RoBERTuito (Spanish Twitter, 500M+ tweets) achieves 0.8692 on Spanish—both substan-

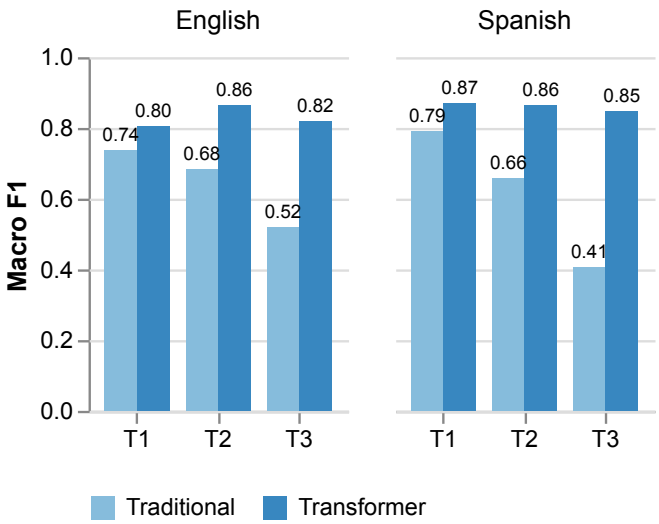

Figure 2: Best traditional ML vs. best transformer validation macro F1 by task and language. Task 3 shows the largest gain, especially in Spanish (0.41 → 0.85).

tially outperforming the multilingual mBERT (0.7649 joint). For Tasks 2 and 3 requiring finer cross-lingual semantic distinctions, XLM-RoBERTa-Large (0.8631 / 0.8195) outperforms all other models, while RoBERTuito also performs well on Spanish (Task 2: 0.8374, Task 3: 0.8455). Figure 2 shows the largest traditional-to-transformer gains on Task 3, where transformer fine-tuning improves English from 0.52 to 0.82 and Spanish from 0.41 to 0.85. Per-class F1 on Task 3 (reported in the Appendix, Table 11) correlates strongly with sample count: Nation (786 samples) achieves 0.86, while Religion (15 samples) drops to 0.72 despite CB-Focal weighting.

4.3. Final Ensemble and Official Test Results

Based on the validation results above, we construct the final ensemble as described in Section 3: a weighted average of BERTweet and mBERT for English Task 1 (weights 0.504 / 0.496), a weighted average of RoBERTuito and Spanish BERT for Spanish Task 1 (0.507 / 0.493), and XLM-RoBERTa-Large for Tasks 2 and 3 (both languages). Table 4 presents the held-out test-set scores.

Table 4
Held-out test-set macro F1 scores. Best result per task highlighted in bold per language.

Lang	Model	Task 1	Task 2	Task 3
EN	BERTweet + mBERT (ensemble)	**0.8291**	–	–
EN	XLM-RoBERTa-Large	–	**0.8474**	**0.8605**
ES	RoBERTuito + Spanish BERT (ensemble)	**0.8707**	–	–
ES	XLM-RoBERTa-Large	–	**0.8978**	**0.9211**
Average EN+ES		**0.8499**	**0.8726**	**0.8908**

The final system achieved an average macro F1 of 0.8711 across all six (language, task) combinations. Complete validation metrics (accuracy, balanced accuracy, precision, recall, F1, weighted metrics, and per-class support) for the final model family are reported in the Appendix (Tables 10 and 11).

4.4. Impact of Class-Balanced Focal Loss

Table 5 and Figure 3 compare Task 3 performance with and without CB-Focal. Without CB-Focal, the best traditional ML model (psycholinguistic XGBoost, using standard cross-entropy via sklearn) achieves 0.5175 macro F1 on English Task 3 and 0.4069 on Spanish. With transformer fine-tuning and

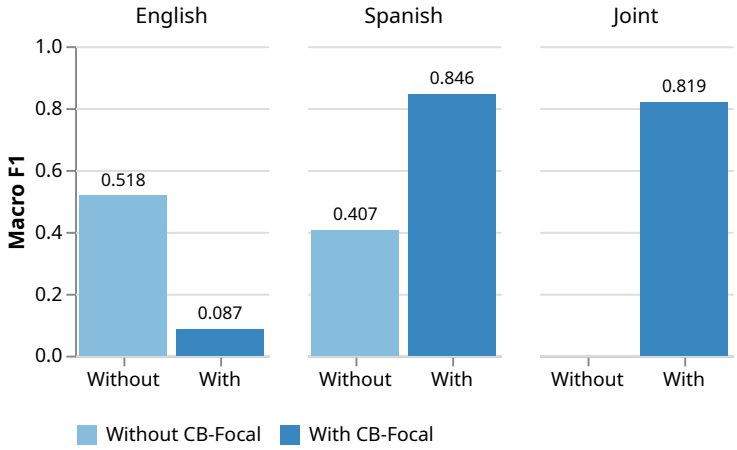

Figure 3: Task 3 macro F1: best traditional ML (no focal loss) vs. best transformer with CB-Focal. The improvement reflects the combined effect of transformer architecture and CB-Focal loss.

CB-Focal (β=0.9, γ=2), performance improves substantially: individual transformer models achieve 0.7994–0.8455, and XLM-RoBERTa-Large reaches 0.8605 (EN) and 0.9211 (ES) on the held-out test set. This improvement reflects the combined effect of two changes: (1) the transformer architecture, which provides richer contextual representations than psycholinguistic feature vectors, and (2) CB-Focal loss, which directly addresses the 52× class imbalance (Nation: 786 samples vs. Religion: 15). While we do not isolate the individual contribution of CB-Focal from the transformer architecture in a controlled ablation, prior work [14] demonstrates that CB-Focal alone provides significant gains over cross-entropy under comparable imbalance ratios, making both components likely contributors to the observed improvement.

Table 5
Task 3 comparison: baseline loss (traditional ML, no focal loss) vs. CB-Focal (transformer). Macro F1 on validation (val) and held-out test set.

Model	Loss	Lang	Val	Test
XGBoost (psycholinguistic)	Cross-entropy	EN	0.5175	—
XGBoost (psycholinguistic)	Cross-entropy	ES	0.4069	—
mBERT (joint)	CB-Focal	EN	0.7994	—
mBERT (joint)	CB-Focal	ES	0.7994	—
RoBERTuito (ES)	CB-Focal	ES	0.8455	—
XLM-RoBERTa-Large (joint)	CB-Focal	EN	0.8195	**0.8605**
XLM-RoBERTa-Large (joint)	CB-Focal	ES	0.8195	**0.9211**

5. Discussion

5.1. Why Psycholinguistic Features Outperform Embeddings

The dominance of 108-dimensional psycholinguistic features over 3000+ dimensional API embeddings suggests that social support expression is characterized by specific affective and cognitive language patterns rather than general semantic content. Categories such as ”positive emotion,” ”social connection,” and ”help” in Empath directly capture the linguistic correlates of supportive behavior. Generic embeddings, trained on broad corpora for semantic similarity, may not adequately emphasize these specific dimensions.

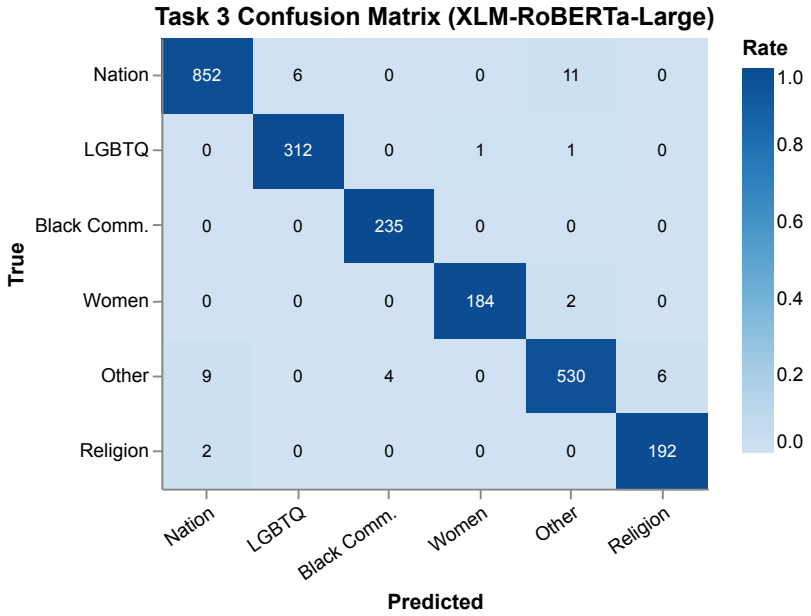

Figure 4: Task 3 confusion matrix (XLM-RoBERTa-Large). Off-diagonal counts reveal confusion between Nation ↔ Other and Women ↔ Black Community.

5.2. Domain Pre-training vs. Multilingual Capacity

Our findings on the task-dependent superiority of domain-specific versus general multilingual models answer RQ3 clearly. For Task 1 (binary support detection), social media pre-training dominates: BERTweet (0.8039 EN val) outperforms mBERT (0.7649) and XLM-RoBERTa-Large (0.7611) on English, and RoBERTuito (0.8692 ES val) substantially outperforms Spanish BERT (0.8437) and XLM-R (0.7611) on Spanish. The advantage is largest for Task 1, where the informal, emoji-rich register of YouTube comments most closely resembles the Twitter pretraining data. For Tasks 2 and 3, which require finer cross-lingual semantic distinctions between group-level support categories, XLM-RoBERTa-Large's larger capacity (560M parameters, 100-language pretraining corpus) yields the best results—suggesting that multilingual breadth and model scale become more valuable as classification granularity increases.

5.3. Late Fusion of Transformers and Traditional ML (RQ4)

We evaluated a late fusion ensemble combining transformer probabilities with psycholinguistic XGBoost scores via grid-searched weights. On the held-out validation set this did not consistently outperform the pure-transformer configuration: gains on Task 1 were marginal (<0.002 macro F1) and Task 3 degraded slightly when XGBoost predictions (capped at 0.52 macro F1) were blended into the stronger XLM-R signal. The final system therefore uses transformer-only predictions for Tasks 2 and 3, and a weighted average of two domain-specific transformers for Task 1 (BERTweet + mBERT for English; RoBERTuito + Spanish BERT for Spanish). This ensemble yields the test scores reported in Table 4: EN Task 1 **0.8291**, ES Task 1 **0.8707**, Tasks 2–3 **0.8474–0.9211**.

5.4. Limitations and Future Work

Our approach has several limitations. API embedding experiments were affected by an indexing bug in the Spanish data (Spanish IDs mapped to English embeddings), though given the consistently poor performance of all embedding sources relative to psycholinguistic features, this is unlikely to change conclusions.

Figure 4 shows the Task 3 confusion matrix for XLM-RoBERTa-Large. The model correctly classifies the majority of Nation and Other examples, but systematic confusion persists between semanti-

cally adjacent groups: Nation↔Other (both relate to national/political identity), and Women↔Black Community (both involve identity-based marginalisation). These confusions are consistent with the overlapping vocabulary in supportive discourse targeting these groups.

LLM augmentation (GLM-4, Claude Sonnet, Gemini 2.0 Flash) improved Task 3 English performance by +0.045 macro F1 (mBERT: 0.790 → 0.835), while Spanish performance changed by −0.017 (0.827 → 0.811). The Spanish regression may reflect that the language instruction added to the prompt—"escribe este comentario en español"—was sufficient for surface fluency but insufficient for matching the specific lexical patterns that characterize Spanish-language social support expressions. Future work should investigate class-conditional prompt design and semantic similarity filtering to improve synthetic sample quality.

The label remapping requirement for Task 3 (non-contiguous labels 0, 1, 2, 4, 5, 6) is a dataset-specific pitfall that would silently produce wrong results without explicit handling. We recommend verifying label contiguity before training multiclass models on any similar dataset.

6. Conclusion

We present a hierarchical transformer ensemble for social support detection achieving strong performance across all three subtasks. Key findings include: (1) psycholinguistic features (108-d) consistently outperform high-dimensional API embeddings (768–4096-d) for this affective classification task, with concatenation actively degrading performance; (2) domain-specific social media pre-training (BERTweet, RoBERTuito) dominates Task 1 over general-purpose multilingual models, while XLM-RoBERTa-Large's larger capacity yields best results on Tasks 2–3; (3) Class-Balanced Focal Loss with β=0.9, combined with transformer fine-tuning, substantially improves Task 3 under severe imbalance—from 0.52 (best traditional ML, no focal loss) to 0.86–0.92 (transformer with CB-Focal) on the held-out test set; and (4) a non-contiguous label space in Task 3 would silently cause 0.37–0.42 macro F1 despite correct training, requiring explicit forward/inverse label remapping. On the held-out test set, we obtained average macro F1 of **0.8499** / **0.8726** / **0.8908** for Tasks 1/2/3 respectively (0.8711 overall average). Code and trained models are available upon acceptance.

Declaration on Generative AI

During the preparation of this work, the authors used the following AI-assisted tools: Claude (Anthropic) for implementation assistance and code debugging; GPT 5.5 (OpenAI) for grammar checking and document revision; GLM 5.1 (Zhipu AI, via Synthetic.new) for code writing assistance; Kimi K 2.6 (Moonshot AI, via Synthetic.new) for visual analysis of figures; and GLM 5.2 (Zhipu AI, via Synthetic.new) for text revision. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. S. House, Work stress and social support, Addison-Wesley, Reading, MA, 1981.

[2] M. Shahiki Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, G. Sidorov, O. Kolesnikova, Online social support detection in Spanish social media texts, arXiv preprint (2025). arXiv:2502.09640.

[3] Z. Ahani, M. Shahiki Tash, F. Balouchzahi, L. Ramos, G. Sidorov, A. Gelbukh, Social support detection from social media texts, PLOS ONE 21 (2026) e0337476. doi:10.1371/journal.pone.0337476.

[4] O. Kolesnikova, M. Shahiki 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.

[5] M. Shahiki Tash, Z. Ahani, L. Ramos, et al., Linguistic analysis in different types of support using LIWC for Spanish and English online communities, Research Square (2025). doi:10.21203/rs.3.rs-9557287/v1.

[6] 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.

[7] 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.

[8] 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. URL: https://aclanthology.org/2020.emnlp-demos.2.

[9] 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. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[10] C. E. Cutrona, D. W. Russell, Type of social support and specific stress: Toward a theory of optimal matching, in: I. G. Sarason, B. R. Sarason, G. R. Pierce (Eds.), Social support: An interactional view, Wiley, New York, 1990, pp. 319–366.

[11] D. Yang, R. E. Kraut, Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019, pp. 1–14. doi:10.1145/3290605.3300574.

[12] C. N. Silla, A. A. Freitas, A survey of hierarchical classification across different application domains, Data Mining and Knowledge Discovery 22 (2011) 31–72. doi:10.1007/s10618-010-0175-9.

[13] 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, 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.

[14] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277. doi:10.1109/CVPR.2019.00949.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, et al., 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. URL: https://aclanthology.org/2020.acl-main.747.

[17] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Practical ML for Developing Countries Workshop at ICLR 2020, 2020. URL: https://arxiv.org/abs/2308.02976.

[18] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493.

[19] E. Fast, B. Chen, M. S. Bernstein, Empath: Understanding topic signals in large-scale text, in:

Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 4647–4657. doi:10.1145/2858036.2858535.

[20] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A Python toolkit for opinion mining and social NLP tasks, arXiv preprint (2021). arXiv:2106.09462.

[21] A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, et al., ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools, arXiv preprint (2024). arXiv:2406.12793.

A. Reproducibility

All experiments use random seed 42. Hyperparameters and model checkpoints are available in the artifacts directory. The pipeline is implemented as numbered stages (01-10) with Parquet intermediate files ensuring reproducible execution.

LLM Augmentation Prompt

The following prompt template was used to generate synthetic Task 3 training examples with each of the three LLMs (GLM-4, Claude Sonnet, Gemini 2.0 Flash). `{class_name}` is the group target label (e.g., "LGBTQ", "Religion"); `{lang_instruction}` is "in English" or "in Spanish"; `{examples}` is three randomly sampled real comments from the training set for that class and language.

```
System: You are a helpful assistant that generates
        realistic YouTube comments.

User: You are generating training data for a social
support detection model.

Write 20 short YouTube comments (1-3 sentences each)
that express support specifically for {class_name}
people. Each comment should sound natural and
conversational, like a real YouTube comment.
Write one comment per line, no numbering.
Write all comments {lang_instruction}.

Examples from the dataset:
{examples}
```

Generation parameters: temperature 0.7, max_tokens 2000. Generated samples are filtered to at most 20 per (class, language, model) combination and appended only to training folds, never to the validation split.

Appendix: Full Traditional ML Results

Complete ablation grids. Tables 6–8 report the full traditional-ML ablation across all feature sources, classifiers, and tasks. Table 9 reports complete validation metrics (accuracy, balanced accuracy, precision, recall, F1) for the strongest traditional setting.

Complete model metrics. Tables 10 and 11 report the validation metrics available for the final model family beyond macro and weighted F1.

Table 6
Task 1 (Support Detection) Traditional ML Results (5-fold CV). Each cell shows Macro F1 ±std (top) / Weighted F1 (bottom). Bold = best per column.

	English				Spanish			
Feature Source	**LR**	**SVC**	**RF**	**XGB**	**LR**	**SVC**	**RF**	**XGB**
Psycholinguistic	0.6708±.014 0.7497	0.6741±.011 0.7503	**0.6994**±.011 0.8093	**0.7354**±.009 0.8134	**0.7892**±.018 0.8504	**0.7906**±.014 0.8527	**0.7846**±.022 0.8572	0.7850±.007 0.8498
TF-IDF	0.5011±.008 0.6593	0.5007±.007 0.6754	0.4426±.003 0.6778	0.4822±.007 0.6819	0.4750±.014 0.6614	0.4773±.010 0.6735	0.4354±.001 0.6732	0.4463±.008 0.6707
Word2Vec	0.4824±.013 0.5724	0.4813±.012 0.5689	0.4531±.007 0.6809	0.4976±.014 0.6738	0.4430±.025 0.5082	0.4514±.030 0.5210	0.4391±.003 0.6719	0.4832±.009 0.6742
Nomic	0.4782±.007 0.5773	0.4794±.008 0.5833	0.4391±.002 0.6775	0.4797±.007 0.6850	0.4693±.019 0.5752	0.4784±.016 0.5880	0.4355±.000 0.6734	0.4445±.011 0.6737
Gemini	0.4882±.008 0.5914	0.4878±.010 0.5955	0.4393±.002 0.6778	0.4606±.010 0.6806	0.4763±.004 0.5838	0.4789±.010 0.5978	0.4355±.000 0.6734	0.4540±.013 0.6796
Qwen3	0.4797±.013 0.5849	0.4720±.010 0.5785	—	0.4563±.007 0.6797	0.4754±.015 0.5874	0.4766±.014 0.5950	—	0.4433±.006 0.6748
OpenAI-Large	0.4886±.016 0.5894	0.4847±.005 0.5885	0.4393±.002 0.6778	0.4598±.010 0.6803	0.4825±.020 0.5901	0.4864±.018 0.5971	0.4355±.000 0.6734	0.4414±.009 0.6736

LR = Logistic Regression, SVC = Linear SVC, RF = Random Forest, XGB = XGBoost. Dashes = not run (timeout/error). Psych = Psycholinguistic features.

Table 7
Task 2 (Target Identification) Traditional ML Results (5-fold CV). Each cell shows Macro F1 ±std (top) / Weighted F1 (bottom). Bold = best per column.

	English				Spanish			
Feature Source	**LR**	**SVC**	**RF**	**XGB**	**LR**	**SVC**	**RF**	**XGB**
Psycholinguistic	**0.6459**±.027 0.7526	**0.6596**±.033 0.7654	**0.5778**±.025 0.7724	**0.6836**±.033 0.8093	**0.6163**±.041 0.6763	**0.6106**±.040 0.6702	**0.5817**±.049 0.7395	**0.6577**±.045 0.7608
TF-IDF	0.4720±.019 0.7159	0.4666±.011 0.7205	0.4456±.001 0.7200	0.4646±.019 0.7224	0.4469±.031 0.6694	0.4354±.015 0.6657	0.4359±.001 0.6739	0.4490±.026 0.6735
Word2Vec	0.4510±.033 0.5740	0.4503±.032 0.5723	0.4579±.010 0.7230	0.4871±.007 0.7183	0.4264±.030 0.5022	0.4456±.041 0.5321	0.4544±.037 0.6758	0.4909±.060 0.6653
Nomic	0.4767±.027 0.6220	0.4876±.022 0.6342	0.4467±.000 0.7216	0.4570±.016 0.7250	0.5075±.035 0.6235	0.5069±.033 0.6348	0.4359±.001 0.6739	0.4684±.023 0.6785
Gemini	0.4971±.029 0.6483	0.4990±.017 0.6614	0.4467±.000 0.7216	0.4522±.012 0.7237	0.5034±.036 0.6160	0.5293±.041 0.6626	0.4359±.001 0.6739	0.4710±.042 0.6859
Qwen3	0.4715±.019 0.6292	0.4939±.023 0.6538	—	0.4481±.005 0.7207	0.5065±.042 0.6218	0.5110±.036 0.6552	—	0.4737±.034 0.6890
OpenAI-Large	0.5058±.033 0.6531	0.5093±.022 0.6631	0.4467±.000 0.7216	0.4455±.001 0.7197	0.4918±.044 0.6171	0.5112±.041 0.6563	0.4359±.001 0.6739	0.4601±.028 0.6787

Table 8
Task 3 (Group Target Classification) Traditional ML Results. Each cell shows Macro F1 ±std (top) / Weighted F1 (bottom) from 5-fold CV. † = ES non-psycholinguistic values shown without ±std due to small sample size. Bold = best per column.

	English				Spanish			
Feature Source	**LR**	**SVC**	**RF**	**XGB**	**LR**	**SVC**	**RF**	**XGB**
Psycholinguistic	**0.3602**$_{\pm.0338}$ 0.4440	**0.4831**$_{\pm.0242}$ 0.6273	**0.4451**$_{\pm.0764}$ 0.6019	**0.5175**$_{\pm.0261}$ 0.6510	**0.2737**$_{\pm.0328}$ 0.3567	**0.2777**$_{\pm.0237}$ 0.4194	**0.3184**$_{\pm.0207}$ 0.4871	**0.4069**$_{\pm.0317}$ 0.5562
TF-IDF	0.1874$_{\pm.0147}$ 0.4396	0.1712$_{\pm.0082}$ 0.4330	0.1224$_{\pm.0042}$ 0.3897	0.1448$_{\pm.0024}$ 0.4112	0.1499†	0.1308†	0.1053†	0.1493†
Word2Vec	0.1050$_{\pm.0280}$ 0.1530	0.1631$_{\pm.0144}$ 0.3869	0.1553$_{\pm.0156}$ 0.4246	0.1709$_{\pm.0145}$ 0.4180	0.1391†	0.1608†	0.1443†	0.1541†
Nomic	0.1639$_{\pm.0203}$ 0.3001	0.1865$_{\pm.0321}$ 0.3925	0.1169$_{\pm.0005}$ 0.3802	0.1541$_{\pm.0096}$ 0.4217	0.1712†	**0.2044**†	0.1053†	0.1784†
Gemini	0.1631$_{\pm.0203}$ 0.3164	0.1860$_{\pm.0171}$ 0.4084	0.1172$_{\pm.0002}$ 0.3811	0.1617$_{\pm.0154}$ 0.4386	0.1550†	0.1811†	0.1053†	0.1537†
Qwen3	0.1642$_{\pm.0092}$ 0.3164	0.1891$_{\pm.0108}$ 0.4038	—	—	0.1551†	0.1640†	—	—
OpenAI-Large	—	—	—	—	0.1840†	**0.2073**†	—	—

All results from 5-fold cross-validation. † = ES non-psycholinguistic Task 3 values shown without ±std due to small ES sample size (590 supportive rows). Dashes = not run (timeout/error or no gain in LR/SVC).

Table 9
Complete validation metrics for the strongest traditional-ML setting. Metrics are recomputed with the same stratified 5-fold CV protocol.

Lang	Task	Feature	Model	N	Acc.	Bal. Acc.	Macro P	Macro R	Macro F1	Weighted P	Weighted R	Weighted F1
EN	Task 1	Psycholinguistic	XGBoost	7998	0.8321	0.6989	0.7730	0.6989	0.7232	0.8193	0.8321	0.8189
ES	Task 1	Psycholinguistic	LinearSVC	2600	0.8523	0.7919	0.7899	0.7919	0.7906	0.8534	0.8523	0.8527
EN	Task 2	Psycholinguistic	XGBoost	1795	0.8256	0.6281	0.7243	0.6281	0.6510	0.8029	0.8256	0.8030
ES	Task 2	Psycholinguistic	XGBoost	590	0.7763	0.6261	0.6813	0.6261	0.6396	0.7601	0.7763	0.7606
EN	Task 3	Psycholinguistic	XGBoost	1795	0.5733	0.4834	0.4791	0.4834	0.4793	0.5726	0.5733	0.5721
ES	Task 3	Psycholinguistic	XGBoost	590	0.4814	0.3445	0.3465	0.3445	0.3401	0.4708	0.4814	0.4713

Table 10
Complete aggregate validation metrics for the final model family. Metrics are reported on the validation split used for model selection.

Task	Model	Acc.	Bal. Acc.	Macro P	Macro R	Macro F1	Weighted P	Weighted R	Weighted F1
EN Task 1	BERTweet	0.8638	0.8033	0.8045	0.8033	0.8039	0.8635	0.8638	0.8636
ES Task 1	RoBERTuito	0.9077	0.8714	0.8671	0.8714	0.8692	0.9083	0.9077	0.9080
Task 2	XLM-R-Large	0.9119	0.8631	0.8631	0.8631	0.8631	0.9119	0.9119	0.9119
Task 3	XLM-R-Large	0.8302	0.8415	0.8052	0.8415	0.8195	0.8309	0.8302	0.8290

Table 11
Per-class validation precision, recall, and F1 for the final model family.

Task / Model	Class	Precision	Recall	F1	Support
EN Task 1 / BERTweet	0 (Non-Supportive)	0.9115	0.9130	0.9122	1241
EN Task 1 / BERTweet	1 (Supportive)	0.6975	0.6936	0.6955	359
ES Task 1 / RoBERTuito	0 (Non Support)	0.9425	0.9378	0.9401	402
ES Task 1 / RoBERTuito	1 (Support)	0.7917	0.8051	0.7983	118
Task 2 / XLM-R-Large	0 (Group)	0.9449	0.9449	0.9449	381
Task 2 / XLM-R-Large	1 (Individual)	0.7812	0.7812	0.7812	96
Task 3 / XLM-R-Large	0 (Black Community)	0.8077	0.9130	0.8571	23
Task 3 / XLM-R-Large	1 (LGBTQ)	0.9524	0.9375	0.9449	64
Task 3 / XLM-R-Large	2 (Nation)	0.8851	0.9222	0.9032	167
Task 3 / XLM-R-Large	3 (No)	0.8000	0.7083	0.7514	96
Task 3 / XLM-R-Large	4 (Other)	0.7263	0.6970	0.7113	99
Task 3 / XLM-R-Large	5 (Religion)	0.6190	0.8667	0.7222	15
Task 3 / XLM-R-Large	6 (Women)	0.8462	0.8462	0.8462	13