Implicit Target Detection in Social Support Texts Using Contextual Embeddings

Javier André Soto Hernández[1], Sara Rocío Miranda Mateos[2]

[1]*Universidad Panamericana,*
Augusto Rodin No. 498, Insurgentes Mixcoac, Benito Juárez,
Mexico City, 03920, Mexico
[2]*Universidad Panamericana,*
Augusto Rodin No. 498, Insurgentes Mixcoac, Benito Juárez,
Mexico City, 03920, Mexico

Abstract

This work presents a hierarchical multilingual pipeline for social support detection in social media texts across three subtasks: support detection, target identification, and targeted group classification. We compare traditional TF-IDF-based classifiers against Transformer-based models including DistilRoBERTa-base and BETO.

The results show that Transformers perform best in context-dependent support detection tasks, while TF-IDF-based approaches remain highly effective under class imbalance conditions. In particular, TF-IDF + Logistic Regression achieved the best performance for Spanish target identification with a Macro-F1 score of 0.81.

Overall, the findings demonstrate that different subtasks benefit from different modeling strategies depending on their lexical and contextual complexity.

Keywords

Social Support Detection, Implicit Target Detection, Multilingual NLP, Text Classification, Contextual Embeddings, DistilRoBERTa-base, BETO, Class Imbalance, Social Media Analysis

1. Introduction

The rapid growth of social media has transformed the way individuals express support, solidarity, and empathy toward different communities. As a result, **social support detection** has emerged as an important task in Natural Language Processing (NLP), with applications in content moderation, public opinion analysis, and community-focused research. In this context, social support refers to linguistic expressions of encouragement, care, admiration, help, or solidarity directed toward an individual or a group. Unlike general sentiment analysis, which focuses on polarity (e.g., positive or negative), social support detection requires identifying **prosocial intent** and understanding the relational nature of the message.

Equally important is the distinction between supportive and **non-supportive** discourse in social media environments. Non-support is not limited to explicitly hostile language; it can also include indifference, mockery, rejection, misinformation, or comments that shift attention away from care and solidarity. In moderation-oriented settings, this distinction becomes operationally critical because systems must decide not only whether language is positive in tone, but whether it actually provides support to a target. This is especially relevant in short, noisy, and context-dependent posts, where supportive intent may be implicit and where harmful expressions can coexist with references to social groups.

Recent studies have demonstrated that both classical machine learning models and transformer-based approaches can effectively identify supportive language in online texts. For example, Tash et al. (2025) [1] analyze social support detection in Spanish social media, highlighting the linguistic and contextual challenges of short, noisy, user-generated content. At a broader level, Ahani et al. (2026) [2] frame social support detection as a robust and generalizable task across social media settings, while Kolesnikova et al.

IberLEF 2026, September 2026, León, Spain
[†]These authors contributed equally.
0237568@up.edu.mx (J. A. S. Hernández); 0244643@up.edu.mx (S. R. M. Mateos)

(2025) [3] report that advanced machine learning pipelines can improve detection quality when models are carefully tuned and compared. However, most existing work focuses primarily on detecting the presence of support, rather than identifying the specific target of that support. This limitation is critical in real-world scenarios, where understanding who is being supported is necessary for meaningful analysis and downstream applications.

To address this problem, the task can be structured into three subtasks of increasing complexity. The first subtask consists of **binary support detection**, where the goal is to determine whether a given comment expresses support (*Support* vs. *Not Support*). The second subtask focuses on **binary target identification**, distinguishing whether the support is directed toward an individual or a group (*Individual* vs. *Group*). Finally, the third subtask, which represents the main challenge of this work, involves **targeted group multiclass classification**, where the system must identify the specific community being supported (e.g., *Women, LGBTQ, Religion*, etc.). This final step is particularly difficult due to the frequent presence of **implicit expressions**, where the target is not explicitly mentioned in the text.

In this work, we propose a pipeline for targeted social support detection that handles English and Spanish data through parallel modeling approaches. Rather than merging both languages into a single representation, we design models that can process each language appropriately while maintaining a consistent task formulation. The pipeline includes label normalization into six target groups, along with a policy-aligned mechanism that maps comments containing explicit harm or violence to a "Not Support" category. This design ensures that the system reflects real-world moderation constraints while maintaining consistency in classification.

We evaluate both a strong classical baseline based on TF-IDF [4] and Logistic Regression with class balancing, and contextual embedding approaches using transformer-based models. Our analysis goes beyond aggregate performance by incorporating class-level metrics, confusion matrices, and diagnostic visualizations to examine how models behave under ambiguity and class imbalance. In particular, we focus on the ability of models to correctly identify **implicit targets**, a setting that is often overlooked in prior work.

The contributions of this paper are threefold. First, we formalize the problem of implicit targeted social support detection in a multilingual context. Second, we present a reproducible and policy-aligned pipeline that combines data preprocessing, modeling, and evaluation within a unified framework. Third, we provide an empirical analysis comparing classical and transformer-based models, highlighting their strengths and limitations in handling ambiguity and minority classes. Our findings underscore the importance of context-aware modeling and careful evaluation when addressing real-world social support detection tasks.

We adopt a two-tier modeling strategy. As language-agnostic baselines, we evaluate classical machine learning approaches — including Logistic Regression, Support Vector Machines, and Random Forest — applied uniformly across both English and Spanish data. For the transformer-based tier, we select language-specific models suited to each corpus: DistilRoBERTa for English, a distilled variant of RoBERTa [5] produced using the same knowledge-distillation approach introduced for DistilBERT [6], retaining a substantial share of its parent model's contextual representational capacity at reduced parameter count and computational cost; and BETO [7] for Spanish, a BERT-based model trained on large Spanish corpora that has demonstrated competitive results across a range of Spanish NLP tasks. For dense vector representations, we complement the classical tier with language-specific embeddings: GloVe [8] for English and FastText [9] for Spanish, the latter selected as a Spanish-language counterpart to GloVe, given its strong coverage of Spanish vocabulary and prior use in Spanish NLP tasks.

2. Related Work

Social support detection in online texts has gained increasing attention as an extension of traditional sentiment analysis, focusing on identifying prosocial and supportive language in social media. Early approaches relied on classical machine learning techniques combined with surface-level textual features,

such as bag-of-words and TF-IDF representations. These methods demonstrated effectiveness in detecting explicit expressions of support but were limited in capturing contextual meaning and implicit references.

Recent work has explored more advanced approaches based on deep learning and transformer architectures. Tash et al. (2025) [1] focus on Spanish-language social media, showing that domain-specific noise and implicit phrasing require more context-aware modeling than standard lexical baselines. Ahani et al. (2026) [2] provide a broader study of social support detection from social media texts and reinforce the relevance of this task for reliable automatic analysis. Kolesnikova et al. (2025) [3] compare advanced machine learning techniques for social support detection and report strong gains from modern models over simpler traditional configurations, especially for subtle semantic patterns.

Relatedly, Tash et al. (2025) [10] apply LIWC-based linguistic analysis to characterize different types of support across Spanish and English online communities, offering a complementary lexical perspective to the classification-oriented approaches above.

Despite these advances, most existing studies primarily address the binary task of detecting whether support is present, without explicitly modeling the target of that support. Only limited attention has been given to distinguishing whether support is directed toward individuals or groups, and even less to identifying the specific community being supported. This gap becomes more pronounced in real-world social media data, where support is often expressed implicitly and without direct mention of the target.

In addition, prior work has largely focused on aggregate performance metrics, with less emphasis on analyzing model behavior under ambiguity or class imbalance. This is particularly relevant in targeted classification tasks, where minority classes and implicit expressions can significantly impact model reliability.

In this work, we extend existing research by addressing the problem of **implicit targeted social support detection** in a multilingual setting. Unlike prior approaches, we focus on identifying the specific group being supported, even when the target is not explicitly mentioned. We also incorporate a policy-aligned labeling strategy that accounts for harmful or violent content, and provide a detailed evaluation that includes class-level analysis and the impact of imbalance. By combining classical baselines with contextual embedding models, this work offers a comprehensive perspective on the strengths and limitations of current approaches for real-world social support detection.

3. Task Definition

This work is developed within the framework of the IberLEF 2026 shared task on social support detection[11][12], which is structured into three sequential subtasks of increasing complexity. The overall objective is to analyze social media comments and determine not only whether support is expressed, but also the nature and target of that support.

3.0.1. Subtask 1: Binary Support Detection

The first subtask consists of **binary support detection**, where the system determines whether a comment expresses support (*Support* vs. *Not Support*). Support is defined as expressions of encouragement, care, admiration, help, or solidarity toward an individual or group.

3.0.2. Subtask 2: Binary Target Identification

The second subtask focuses on **binary target identification**, where the system determines whether the support is directed toward an individual or a group (*Individual* vs. *Group*).

3.0.3. Subtask 3: Targeted Group Classification

The primary focus of this work is **Subtask 3**, which consists of targeted group multiclass classification. Given a single supportive comment aimed at a group, the goal is to identify which specific community

is being supported. The possible output classes are: *Nation, Other, LGBTQ, Black Community, Religion,* and *Women*.

In addition to the classification objective, the task includes a scope constraint stating that any comment explicitly promoting harm or violence must be labeled as *Not Support*, regardless of the target being referenced. This constraint introduces an additional level of complexity, requiring the system to distinguish between supportive intent and harmful content.

A key challenge of Subtask 3 is the presence of **implicit target expressions**, where the supported group is not explicitly mentioned in the text. Instead, the target must be inferred from contextual cues, making this task significantly more complex than traditional text classification problems.

3.1. Dataset Description and Class Distributions

To characterize the data underlying each subtask, we report class distributions for the training population at each stage of the pipeline. Non-eligible instances are omitted from Subtasks 2 and 3 distributions, as they reflect upstream filtering rather than task-relevant class structure.

3.1.1. Subtask 1: Support vs. Not Support

Figure 1 shows the distribution of the two Subtask 1 classes. The English subset (n=7998) consists of 6203 Not Support (77.6%) and 1795 Support (22.4%) instances; the Spanish subset (n=2600) shows a nearly identical proportional split, with 2010 Not Support (77.3%) and 590 Support (22.7%) instances. Despite substantial differences in topical content between the two languages (Section 3.1.4), the degree of class imbalance at this stage is consistent across both.

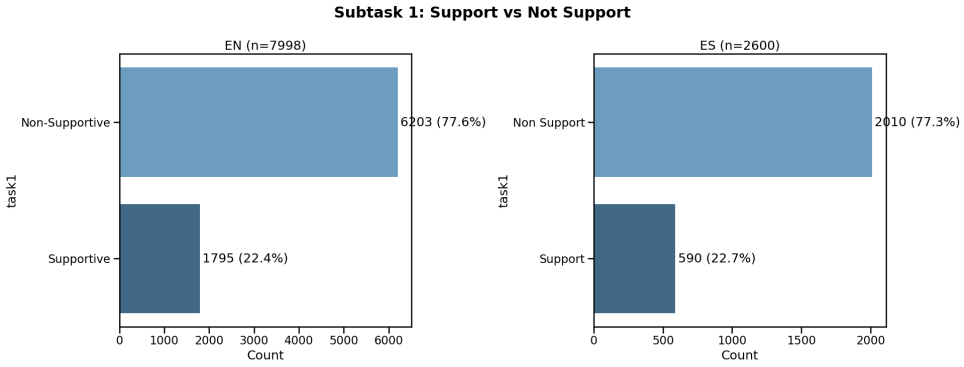

Figure 1: Class distribution for Subtask 1 (Support vs. Not Support), English and Spanish.

3.1.2. Subtask 2: Individual vs. Group

Figure 2 shows the distribution of Individual and Group labels among comments classified as supportive. Both languages show a comparable skew toward Group-directed support: 1450 Group (80.8%) vs. 345 Individual (19.2%) in English, and 456 Group (77.3%) vs. 134 Individual (22.7%) in Spanish.

3.1.3. Subtask 3: Targeted Group Classification

Figure 3 shows the distribution of the six target-group classes among comments classified as both supportive and group-directed. The two languages exhibit markedly different dominant classes: in English, Nation (786, 54.2%) and Other (416, 28.7%) together account for over 80% of instances, while LGBTQ (123, 8.5%), Black Community (91, 6.3%), Women (19, 1.3%), and Religion (15, 1.0%) are comparatively scarce. In Spanish, LGBTQ is instead the dominant class (196, 43.0%), followed by Other (81, 17.8%), Religion (60, 13.2%), Women (48, 10.5%), and Nation (47, 10.3%), with Black Community the smallest class (24, 5.3%). In both languages, the smallest classes contain fewer than 25 instances, which is consistent with the reduced Macro-F1 scores observed for Subtask 3 in Table 5 relative to Weighted-F1 and Accuracy.

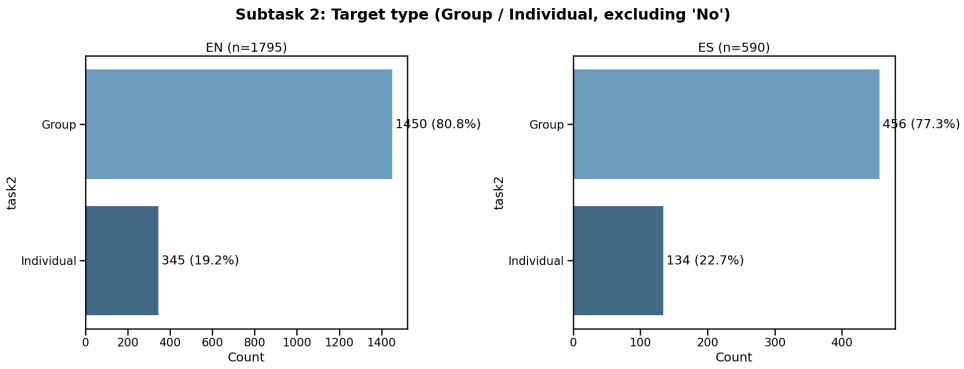

Figure 2: Class distribution for Subtask 2 (Individual vs. Group), among comments classified as supportive.

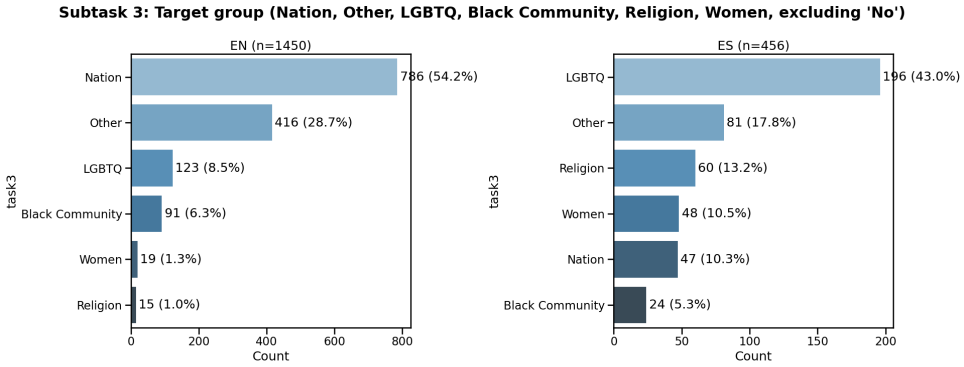

Figure 3: Class distribution for Subtask 3 (target group classification), among comments classified as supportive and group-directed.

3.1.4. Topical Concentration (English)

To assess the geopolitical and topical concentration discussed in Section 6, we examine raw term frequencies in the English subset. Figure 4 shows "palestin" (and variants) appearing in 40.6% of Support instances versus 4.8% of Not Support, with comparable skew for "israel" (12.8% vs. 5.0%), "gaza" (5.2% vs. 1.4%), and "muslim" (5.6% vs. 1.4%).

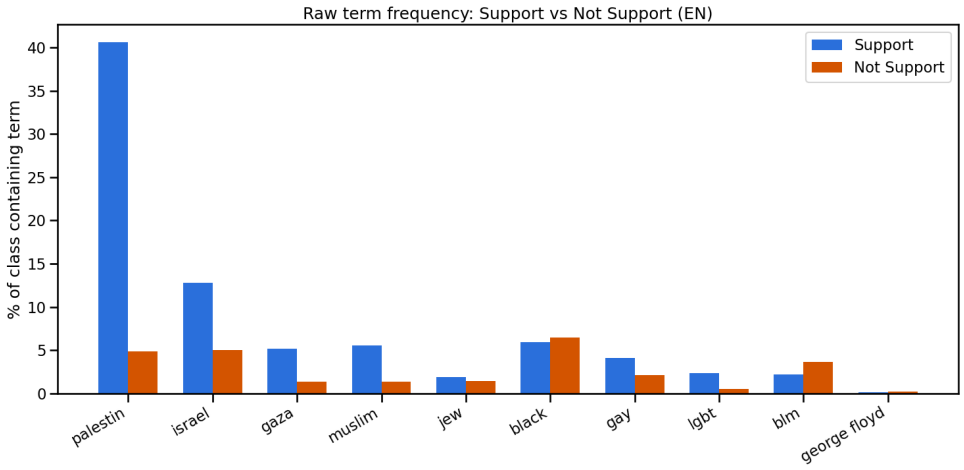

Figure 4: Frequency of selected topical terms across Support and Not Support classes (English).

Figure 5 extends this using a log-odds-ratio approach with an informative Dirichlet prior [13]. Support-associated terms are dominated by vocabulary tied to a single recurring geopolitical conflict

(palestine, israel, muslim, allah) alongside supportive language (love, bless, pray, protect), while Not Support terms are generic and conversational (said, women, money, lol, racist) with no comparable topical concentration. This suggests classifiers may rely on topic-specific vocabulary as a proxy for the Support label rather than generalizable supportive intent — consistent with the LR feature weights in Section 5.1.1 and the class-weighting limitations discussed in Section 6.

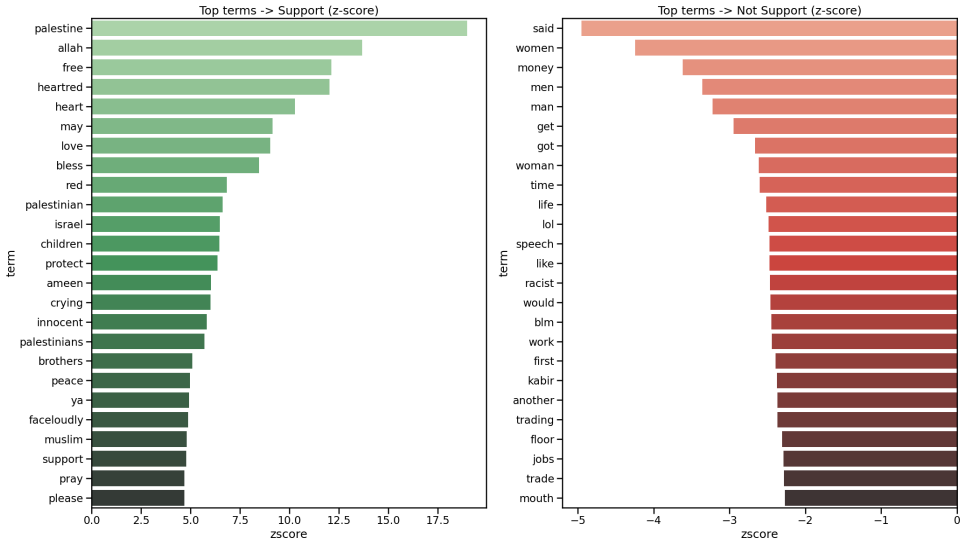

Figure 5: Log-odds ratio with informative Dirichlet prior, English Subtask 1 (Support vs. Not Support).

3.1.5. Topical Concentration (Spanish)

We apply the same analysis to the Spanish subset. Unlike the English data, where topical concentration centers on a single geopolitical conflict, Spanish supportive comments cluster around a specific national sporting or celebratory event involving LGBTQ representation. Figure 6 shows "orgullo" appearing in 18.7% of Support instances versus 1.6% of Not Support, a far stronger skew than any individual LGBTQ-identity term (gay, lgbt, trans, bisexual) shows on its own.

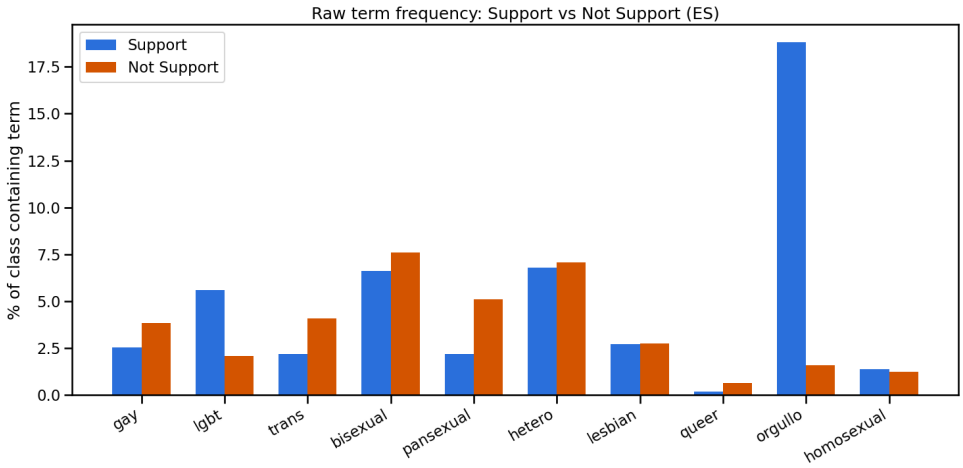

Figure 6: Frequency of selected topical terms across Support and Not Support classes (Spanish).

Figure 7 extends this using the same log-odds-ratio method with an informative Dirichlet prior [13]. Support-associated terms combine national identity (méxico, mexicano, nuestro) with celebratory and sports vocabulary (felicidades, orgullo, viva, atletas, medalla, oro, himno) rather than general LGBTQ-supportive language. Not Support terms are comparatively generic identity-discussion markers

(género, mujer, religión, masculino, binario), with no comparable topical concentration. As in English, this suggests reliance on narrow, event-specific vocabulary as a proxy for supportive intent.

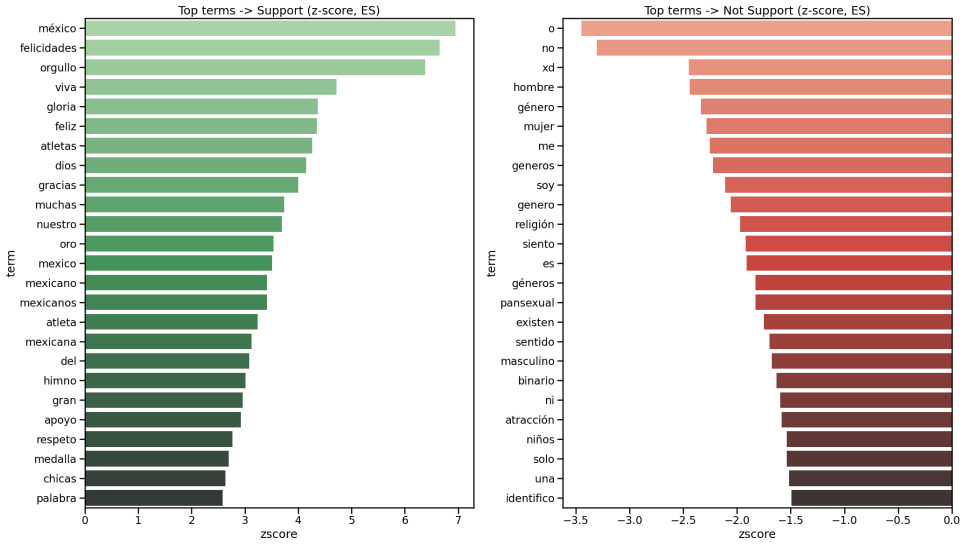

Figure 7: Log-odds ratio with informative Dirichlet prior, Spanish Subtask 1 (Support vs. Not Support).

4. Methodology

This section describes the implementation of our approach across the three subtasks, including data preprocessing, model design, and evaluation strategy. Our methodology follows a hierarchical pipeline, where the outputs of Subtask 1 and Subtask 2 are used to define the input space for Subtask 3.

4.1. Overview of the Pipeline

Our approach consists of three sequential stages:

- Subtask 1: Binary Support Detection
- Subtask 2: Binary Target Identification
- Subtask 3: Targeted Group Classification

Each subtask is implemented as an independent supervised classification problem for English and Spanish datasets. The construction of each subtask's training population and the propagation of labels between stages are described in detail in Section 4.6.

4.2. Subtask 1.EN: English Binary Support Detection

4.2.1. Data Preparation and Preprocessing

For the English binary support detection task, we model the problem as a supervised classification setting with labels *Support* and *Not Support*. Before training, we apply a normalization pipeline to reduce social-media noise and improve generalization. This pipeline includes orthographic correction with SymSpell[14], WordSegment and spaCy-based[15] entity masking, where explicit target mentions are replaced with generic `[SUBJECT]` tokens. Noise in this context refers to three related but distinct phenomena: (a) topical over-concentration around a small number of recurring events, (b) deficient grammatical and syntactic structure, difficult to parse correctly even under manual review (e.g., "*give face plant chp cruiser solid 7 form good follow*"), and (c) typos, unsegmented words, and non-standard symbols. SymSpell and word segmentation address (c), and entity masking addresses (a) by discouraging topic

memorization; (b) was not directly targeted by any preprocessing step and is discussed as a limitation in Section 6. The objective of this masking step is to encourage intent-focused learning rather than topic memorization.

4.2.2. Modeling Strategy

We implement two complementary representation families:

- Sparse lexical representations using TF-IDF features.
- Dense semantic representations using GloVe [8] embeddings.

TF-IDF features were extracted using unigrams, bigrams, and trigrams (`ngram_range=(1,3)`), with minimum and maximum document frequency thresholds (`min_df=5`, `max_df=0.7`) applied to exclude rare outlier terms and overly common topical terms, respectively, and sublinear term-frequency scaling enabled. GloVe-based representations were obtained by averaging 100-dimensional pre-trained word vectors (glove.6B) at the document level; tokens absent from the embedding vocabulary were excluded from the average, and entries with no matched tokens were represented as zero vectors.

On top of these representations, we train three standard classifiers: Logistic Regression, Support Vector Classifier, and Random Forest. Table 1 summarizes the hyperparameter configuration used for these classifiers, which is applied uniformly across all subtasks and languages unless otherwise noted. In parallel, we fine-tune DistilRoBERTa [6] (`DistilRoBERTa-base`), a distilled variant of RoBERTa [5] that retains a substantial share of its contextual representational capacity at reduced parameter count and computational cost, to capture contextual dependencies and implicit support patterns that may be difficult to represent with fixed lexical features. Table 2 summarizes the fine-tuning configuration used for transformer-based models throughout this work.

Table 1
Classical classifier hyperparameter configuration (applied uniformly across all subtasks and languages).

Component	Configuration
TF-IDF Vectorizer	`ngram_range=(1,3)`, `min_df=5`, `max_df=0.7`, `sublinear_tf=True`
Logistic Regression	`class_weight='balanced'`; remaining parameters at default
Random Forest	`class_weight='balanced'`, `max_depth=5`; remaining parameters at default
SVC	`class_weight='balanced'`, `C=0.1`, `kernel='sigmoid'`

A sigmoid kernel was selected for the SVC after preliminary comparison against a linear kernel, which underperformed on the sparse TF-IDF representation in our setting; this configuration was then applied uniformly across subtasks for consistency.

Table 2
Transformer fine-tuning configuration (DistilRoBERTa for English; the same configuration was applied to BETO for Spanish, with the corresponding pre-trained checkpoint and tokenizer).

Parameter	Value
Base checkpoint	DistilRoBERTa-base
Max sequence length	128 (padding to max length, with truncation)
Learning rate	2e-5
Batch size (train/eval)	16 / 16
Epochs	3
Weight decay	0.01
Optimizer	AdamW (Hugging Face `Trainer` default)
LR scheduler	Linear, no warmup (`Trainer` default)
Checkpoint selection	Best epoch by evaluation loss (`load_best_model_at_end=True`)
Train/eval split	80/20, `random_state=42`

4.2.3. Pipeline Output

The output of this stage is a binary support prediction for each English comment. The role of this output in constructing subsequent training populations and in inference-time label propagation is described in Section 4.6.

4.3. Subtask 1.ES: Spanish Binary Support Detection

4.3.1. Approach and Model Selection

For the Spanish support detection subtask, we follow the same binary formulation (*Support* vs. *Not Support*) with language-specific preprocessing. We combine classical feature-based models and transformer-based models to preserve comparability with the English pipeline while accounting for Spanish linguistic characteristics.

This pipeline included:

- **Demojization:** Converting emojis into Spanish textual descriptions (e.g., a heart emoji into "corazón rojo") to incorporate their emotional weight into the transformer's attention mechanism.
- **Full Character Support:** Ensuring the correct identification of unique Spanish characters, such as tildes and the letter "ñ" (áéíóúüñ).
- **SymSpell Orthographic Correction:** Applying a spell-checker trained on Spanish frequency dictionaries to mitigate the noise common in social media text.
- **Refined Tokenization:** Unlike the traditional models, we opted against aggressive lemmatization. Instead, we preserved original word forms (`token.text.lower()`) to allow the model to leverage morphological inflections and contextual nuances.

4.3.2. Modeling Strategy

For classical modeling, we train classifiers on two complementary feature representations: TF-IDF sparse vectors and pre-trained FastText [9] word embeddings (wiki.es, top 500,000 vocabulary by frequency), loaded as static vectors and averaged at the document level following the same procedure used for GloVe in Section 4.2; tokens absent from the vocabulary were excluded from the average, and entries with no matched tokens were represented as zero vectors. Both representations are evaluated with Logistic Regression, Support Vector Classifier, and Random Forest, using the same hyperparameter configuration described in Section 4.2 (Table 1). For contextual modeling, we fine-tune BETO [7] (dccuchile/bert-base-spanish-wwm-uncased), which is pre-trained on Spanish corpora and is therefore suitable for capturing morphosyntactic variation, colloquial writing, and implicit supportive intent in Spanish social media.

4.3.3. Pipeline Output

The output of this stage is a binary support prediction for each Spanish comment. These predictions are propagated to the next stage to constrain target-identification decisions and maintain consistency in the hierarchical pipeline.

4.4. Subtask 2: Individual or Group Classification

4.4.1. Input Definition and Filtering

The primary input to Subtask 2 consists of comments marked as supportive in the training dataset. This design preserves the intended task hierarchy and reduces non-supportive noise before target identification. To maintain consistency across subtasks and avoid distribution shifts, we preserve the same textual normalization pipeline used previously. Consequently, preprocessing operations such as spell correction and word segmentation are inherited directly from the first stage.

4.4.2. Modeling Strategy

We apply the same families of models explored in Subtask 1 — TF-IDF-based classical classifiers, dense embedding models (GloVe for English, FastText for Spanish), and transformer-based architectures (DistilRoBERTa-base for English, BETO for Spanish) — using the hyperparameter configuration described in Section 4.2 (Table 1). This enables a controlled comparison between lexical and contextual representations while maintaining a stable pipeline design across subtasks.

4.4.3. Pipeline Output

The output of Subtask 2 is a binary target label (Individual or Group) assigned to each supportive comment. The role of this output in constructing Subtask 3's training population and in inference-time label propagation is described in Section 4.6.

4.5. Subtask 3: Targeted Group Classification

Subtask 3 is the main focus of this work. The input to this stage consists only of comments that have been classified as *Support* in Subtask 1 and as *Group* in Subtask 2.

4.5.1. Data Preparation and Filtering

The dataset is filtered to include only relevant instances that satisfy the conditions of Subtask 3. Additionally, only samples belonging to the predefined target classes (*Nation, Other, LGBTQ, Black Community, Religion, Women*) are retained. Text normalization is applied to ensure consistent formatting, and invalid or empty entries are removed.

4.5.2. Policy Constraint Handling

To incorporate the task constraint regarding harmful content, we implement a rule-based mechanism that detects explicit violence or harmful language using predefined lexical patterns. These include terms such as *kill, attack, hurt* in English and *matar, golpear, dañar* in Spanish. When such patterns are detected, the corresponding instance is treated as *Not Support*, ensuring alignment with the task requirements.

4.5.3. Language-Specific Pipelines

We design separate pipelines for English and Spanish data rather than merging both languages into a single model. This allows each model to better capture language-specific patterns and linguistic nuances. Each dataset is independently processed, trained, and evaluated.

4.5.4. Modeling Approaches

We evaluate both classical and transformer-based models:

Classical Models TF-IDF representations are used with the following classifiers:

- Logistic Regression
- Linear Support Vector Machine (LinearSVC)
- Random Forest

Unigrams, bigrams and trigrams are used as features, and class imbalance is handled using class weighting strategies.

Transformer-Based Models We fine-tune transformer models for sequence classification:

- *DistilRoBERTa-base* for English
- *dccuchile/bert-base-spanish-wwm-cased* for Spanish

For Subtask 3, we use the cased variant of BETO (dccuchile/bert-base-spanish-wwm-cased) rather than the uncased variant used in Subtask 1, as case information provides additional signal for the entity-sensitive nature of target-group classification, where proper nouns and group/nationality references are more directly relevant.

Tokenization is applied with truncation and a maximum sequence length of 128 tokens. Training is performed for multiple epochs using standard optimization settings.

4.6. Pipeline Construction and Label Propagation

Each subtask is implemented as an independent supervised classification problem for English and Spanish datasets. For each subtask, the training population is constructed by filtering the released training dataset using the gold (ground-truth) labels of the preceding subtask — e.g., Subtask 2's training set consists of all instances where the true Subtask 1 label is *Support*, and Subtask 3's training set consists of all instances where the true Subtask 1 label is *Support* and the true Subtask 2 label is *Group*. This ensures that classifier training is not affected by upstream prediction errors.

At inference time, by contrast, gold labels are unavailable, and the pipeline therefore propagates *predicted* labels between stages: a comment is passed to Subtask 2 only if Subtask 1 predicts *Support*, and to Subtask 3 only if Subtask 2 predicts *Group*. This means that, although training populations are clean, inference-time errors from earlier stages can still constrain which instances reach later stages, a limitation discussed further in Section 6.

4.7. Train-Validation Split

The data is split into training and validation sets using an 80/20 split with stratified sampling to preserve class distribution.

4.8. Evaluation Metrics

Model performance is evaluated using:

- Accuracy
- Macro-F1
- Weighted-F1

Macro-F1 is used as the primary metric for model selection due to its robustness to class imbalance.

4.9. Analysis and Model Selection

To analyze model behavior, we compute confusion matrices and per-class performance metrics. These analyses help identify patterns of misclassification, particularly in minority classes and ambiguous cases.

Models are compared based on their Macro-F1 scores, and the best-performing model is selected accordingly.

5. Results

All results reported in this paper are derived from an internal 80/20 stratified split of the released training dataset, rather than from the official shared-task test set or leaderboard.

5.1. Subtask 1: Binary Support Detection

For Subtask 1, we compared transformer-based and classical models on both languages, selecting the best-performing configuration in each case for use in the subsequent pipeline stages. Performance for all models was measured using a combination of Macro-F1, Weighted-F1, and Accuracy.

5.1.1. English (EN)

Table 3
Subtask 1 model comparison for English.

Model	Macro-F1	Weighted-F1	Accuracy
Transformer (DistilRoBERTa-base)	0.78	0.86	0.86
TF-IDF + SVC	0.76	0.84	0.84
TF-IDF + Logistic Regression	0.76	0.84	0.83
TF-IDF + Random Forest	0.73	0.82	0.82
GloVe + Random Forest	0.71	0.80	0.79
GloVe + Logistic Regression	0.71	0.79	0.78
GloVe + SVC	0.70	0.79	0.78

The best performing model was DistilRoBERTa-base, achieving a Weighted-F1 of 0.86, Macro-F1 of 0.78, and Accuracy of 0.86. This result is consistent with the strengths of transformer-based architectures, which classify support by capturing contextual and semantic relationships between words rather than relying solely on isolated lexical features.

It is also important to note that the dataset was heavily centered around a single geopolitical topic, which introduced significant contextual bias. Terms such as “Palestine” frequently appeared across both supportive and non-supportive samples, causing traditional models to associate topical vocabulary with target labels. Although preprocessing techniques were implemented to reduce this noise and minimize topic memorization, the removal of contextual entities did not lead to a substantial performance improvement. This suggests that part of the classification challenge was intrinsically tied to the highly topic-specific nature of the dataset.

Figures 8 and 9 provide per-class detail absent from aggregate metrics alone. Despite its superior Macro-F1 and Accuracy, DistilRoBERTa achieved a Support-class recall of 0.63 — lower than TF-IDF + LR (0.71) and TF-IDF + SVC (0.68). This indicates that the transformer’s aggregate advantage derives primarily from stronger performance on the majority Not-Support class, while several classical models recover a larger share of true Support instances despite lower overall scores.

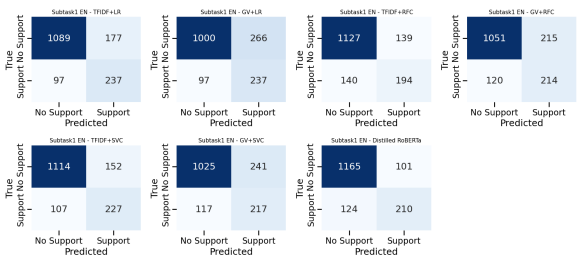

Figure 8: Confusion matrices, Subtask 1 (English).

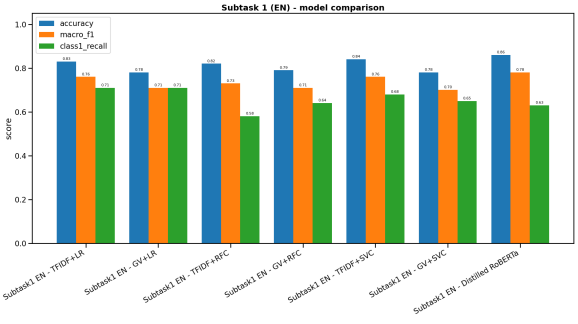

Figure 9: Model comparison, Subtask 1 (English).

5.1.2. Spanish (ES)

The best performing model was BETO-base, achieving a Weighted-F1 of 0.88, Macro-F1 of 0.84, and Accuracy of 0.89. While traditional approaches produced competitive results, simpler

Table 4
Subtask 1 model comparison for Spanish.

Model	Macro-F1	Weighted-F1	Accuracy
Transformer (BETO-base)	0.84	0.88	0.89
TF-IDF + Logistic Regression	0.76	0.83	0.83
FastText + Random Forest	0.75	0.82	0.83
FastText + Logistic Regression	0.75	0.81	0.81
TF-IDF + SVC	0.74	0.82	0.83
FastText + SVC	0.72	0.81	0.81
TF-IDF + Random Forest	0.70	0.79	0.81

non-neural models frequently struggled to correctly identify genuine support, particularly in more context-dependent or ambiguous cases.

The transformer-based architecture demonstrated a stronger ability to capture semantic relationships and contextual meaning, resulting in more consistent classification performance across both classes. Figures 10 and 11 confirm this advantage extends to per-class performance: BETO achieves the highest Support-class recall (0.72), corresponding to 89 of 124 true Support instances correctly identified — a substantially better minority-class recovery rate than any classical model, and a contrast with the English results where no single model dominated on both aggregate and minority-class metrics simultaneously.

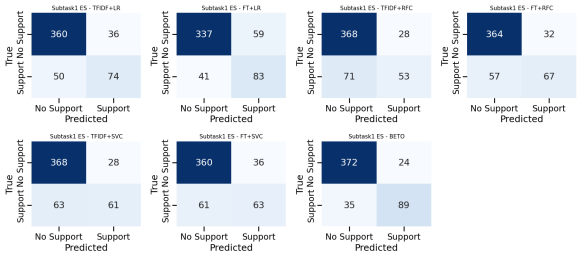

Figure 10: Confusion matrices, Subtask 1 (Spanish).

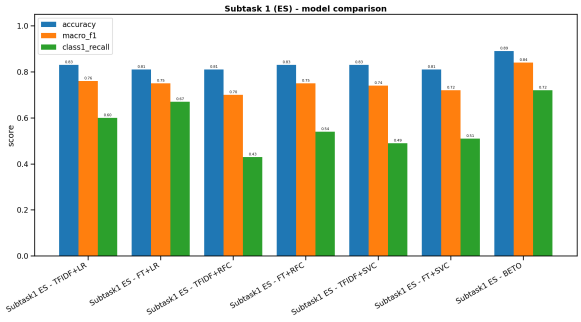

Figure 11: Model comparison, Subtask 1 (Spanish).

5.2. Subtask 2: Binary Target Identification

For group identification, the same model architectures were retrained using the updated target labels. This task introduced additional challenges related to linguistic ambiguity, emoji normalization, and effective word vectorization, particularly due to the informal and inconsistent writing style commonly found in social media data.

5.2.1. Subtask 2 EN

The best performing model was DistilRoBERTa-base, achieving a Weighted-F1 of 0.86, Macro-F1 of 0.74, and Accuracy of 0.86. Compared to Subtask 1, this task introduced a more structurally complex challenge, requiring models to distinguish whether support was directed toward an *Individual* or a *Group* — a distinction that depends on contextual understanding of pronouns and referential expressions rather than topic-level vocabulary.

Traditional TF-IDF approaches performed competitively, with TF-IDF + LR matching DistilRoBERTa's Macro-F1 exactly (0.74). GloVe-based approaches generally underperformed, likely because document-level embedding averaging diluted the structural cues most relevant to this task. Figures 12 and 13 provide per-class context: DistilRoBERTa achieved a Group-class recall of 0.94 (279 of 297 Group instances correctly identified), the highest per-class recall figure observed across any model in any

Table 5
Subtask 2 model comparison for English.

Model	Macro-F1	Weighted-F1	Accuracy
Transformer (DistilRoBERTa-base)	0.74	0.86	0.86
TF-IDF + Logistic Regression	0.74	0.84	0.82
TF-IDF + SVC	0.71	0.81	0.79
TF-IDF + Random Forest	0.70	0.80	0.79
GloVe + Logistic Regression	0.70	0.80	0.77
GloVe + Random Forest	0.69	0.81	0.80
GloVe + SVC	0.61	0.70	0.66

subtask in this study. This strong Group-class identification, combined with competitive Individual-class performance, underpins its aggregate advantage on this task.

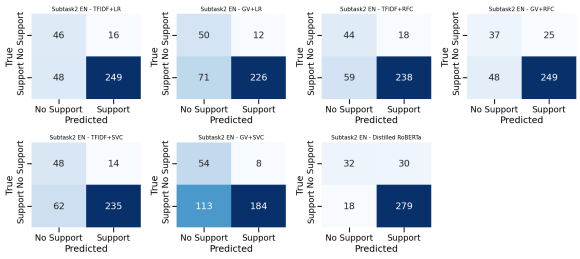

Figure 12: Confusion matrices, Subtask 2 (English).

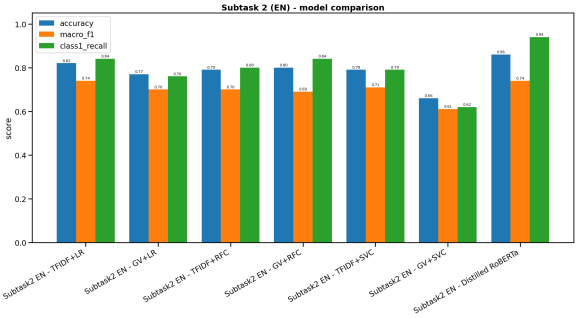

Figure 13: Model comparison, Subtask 2 (English).

5.2.2. Subtask 2 ES

Table 6
Subtask 2 model comparison for Spanish.

Model	Macro-F1	Weighted-F1	Accuracy
TF-IDF + Logistic Regression	0.81	0.87	0.86
TF-IDF + SVC	0.80	0.86	0.86
TF-IDF + Random Forest	0.79	0.86	0.86
FastText + Logistic Regression	0.78	0.84	0.83
Transformer (BETO-base)	0.76	0.84	0.85
FastText + Random Forest	0.70	0.81	0.83
FastText + SVC	0.56	0.60	0.57

TF-IDF-based models achieved the strongest overall performance. TF-IDF + Logistic Regression obtained the highest Macro-F1 (0.81), Weighted-F1 (0.87), and Accuracy (0.86), making it the best-performing model overall. These results suggest that Individual vs. Group classification in Spanish relies heavily on explicit lexical and syntactic markers — pronouns, collective nouns, sentence structure — rather than deep contextual semantics, making sparse lexical representations particularly effective here.

FastText-based approaches produced weaker and less stable results. Figures 14 and 15 make the extreme case visible: FastText + SVC produced a near-degenerate confusion matrix (43 true negatives vs. 48 false positives, 3 false negatives vs. 24 true positives), indicating the model was predicting the majority class for the overwhelming majority of instances rather than learning a discriminative decision boundary. This explains its 0.56 Macro-F1 despite nominal accuracy of 0.57.

Although BETO achieved a Weighted-F1 of 0.84 and Accuracy of 0.85, TF-IDF + Logistic Regression maintained superior classification balance as reflected in its higher Macro-F1. The transformer's higher computational cost was not justified by a measurable improvement in evaluation metrics for this particular task, and TF-IDF + Logistic Regression was therefore selected as the final model for Subtask 2 in Spanish.

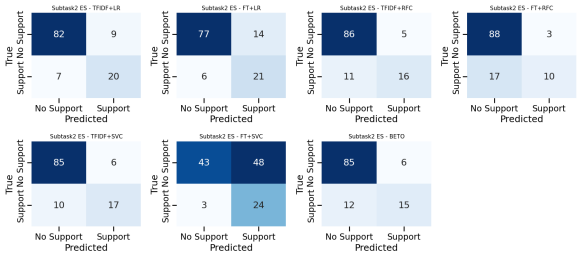

Figure 14: Confusion matrices, Subtask 2 (Spanish).

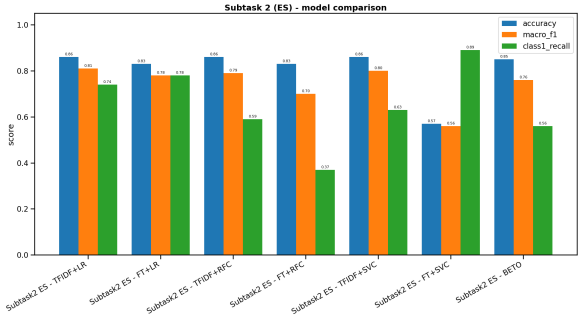

Figure 15: Model comparison, Subtask 2 (Spanish).

5.3. Subtask 3: Targeted Group Classification

For Subtask 3, we compared four model families in each language-specific pipeline: TF-IDF + Logistic Regression, TF-IDF + LinearSVC, TF-IDF + Random Forest, and a transformer baseline.

Table 7
Subtask 3 model comparison for English and Spanish.

Language	Model	Macro-F1	Weighted-F1	Accuracy
EN	TF-IDF + Logistic Regression	0.6955	0.7981	0.8000
EN	TF-IDF + LinearSVC	0.6940	0.7977	0.8000
EN	TF-IDF + Random Forest	0.5285	0.7687	0.7759
EN	Transformer (DistilRoBERTa-base)	0.5722	0.8491	0.8621
ES	TF-IDF + Logistic Regression	0.7081	0.7658	0.7717
ES	TF-IDF + LinearSVC	0.7188	0.7762	0.7826
ES	TF-IDF + Random Forest	0.5938	0.6809	0.6957
ES	Transformer (BETO-cased)	0.7014	0.7931	0.8043

English (EN). In the English setting, the best model according to Macro-F1 was **TF-IDF + Logistic Regression** with **Macro-F1 = 0.6955**, **Weighted-F1 = 0.7981**, and **Accuracy = 0.8000**. TF-IDF + LinearSVC obtained a very similar Macro-F1 (0.6940), while the transformer model (*DistilRoBERTa-base*) achieved higher weighted performance and accuracy (0.8491 and 0.8621, respectively) but lower Macro-F1 (0.5722).

Spanish (ES). In the Spanish setting, the best model according to Macro-F1 was **TF-IDF + LinearSVC** with **Macro-F1 = 0.7188**, **Weighted-F1 = 0.7762**, and **Accuracy = 0.7826**. TF-IDF + Logistic Regression remained competitive (Macro-F1 = 0.7081). The transformer model (*dccuchile/bert-base-spanish-wwm-cased*) reached the highest accuracy (0.8043) and weighted performance (0.7931), but not the highest Macro-F1 (0.7014).

Summary. Using Macro-F1 as the primary selection criterion for class-imbalance robustness, the selected Subtask 3 models were TF-IDF + Logistic Regression for English and TF-IDF + LinearSVC for Spanish. Figures 16 and 17 illustrate the per-class classification behavior and aggregate metric

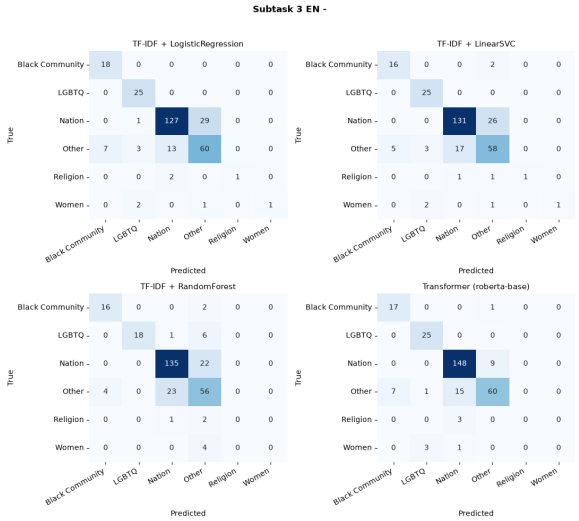

Figure 16: Confusion matrices, Subtask 3 (Spanish).

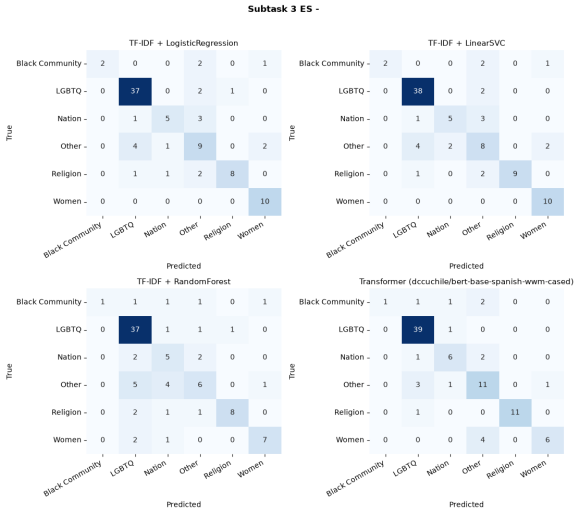

Figure 17: Model comparison, Subtask 3 (Spanish).

comparison for the Spanish pipeline, confirming that minority classes such as Black Community and Women remain the primary source of misclassification across all models.

6. Discussion

The experimental results revealed several important limitations and trade-offs across the evaluated architectures and subtasks. In Subtask 3, Transformer-based models achieved the highest Accuracy and Weighted-F1 scores in both languages; however, they consistently underperformed in Macro-F1 when compared to traditional lexical classifiers. In English, for example, DistilRoBERTa-base obtained a Macro-F1 of 0.5688, substantially below the 0.6955 achieved by TF-IDF + Logistic Regression. This suggests that class-weighting strategies in deep learning were insufficient to compensate for severe class imbalance, particularly for minority target groups.

We also note that Random Forest consistently underperformed relative to the other classical classifiers across every subtask and language, despite empirical tuning. Preliminary experiments without depth constraints showed even weaker generalization, and *max depth=5* was selected empirically to mitigate this. We attribute the residual gap to the nature of TF-IDF representations themselves: high-dimensional, sparse, and largely orthogonal lexical features are poorly suited to the axis-aligned, recursive partitioning that tree-based ensembles rely on, compared to the linear decision boundaries learned by Logistic Regression and SVC.

The preprocessing pipeline, which incorporated SymSpell normalization and entity masking through the `[SUBJECT]` token, partially reduced topic memorization but did not produce substantial improvements in overall performance. **This indicates that the dataset itself remained strongly biased toward a narrow geopolitical context, where recurrent topical entities influenced classification behavior independently of supportive intent.**

An additional finding emerged in Subtask 2 for Spanish, where TF-IDF-based approaches consistently outperformed BETO-base. TF-IDF + Logistic Regression achieved the highest Macro-F1 score of 0.81, compared to 0.76 for BETO. This suggests that the identification of whether support targets an individual or a group depends primarily on explicit lexical and syntactic markers, such as pronouns and collective nouns, rather than deep contextual semantics. Consequently, the computational cost of Transformer-based architectures was not justified for this particular classification problem.

Finally, although the original proposal considered multilingual contextual embeddings, the final methodology relied on specialized monolingual architectures for each language.

The use of DistilRoBERTa-base for English and BETO for Spanish provided stronger adaptation to language-specific morphosyntactic patterns and colloquial social media expressions. **These results suggest that specialized monolingual representations may remain more effective than unified multilingual spaces for highly domain-specific social discourse analysis.**

7. Conclusion

This work explored the effectiveness of traditional machine learning and Transformer-based architectures for multilingual social support detection across three subtasks involving support identification, target classification, and group categorization. **The results demonstrated that no single architecture consistently outperformed all others across every task, highlighting the importance of selecting models according to the linguistic and structural characteristics of each classification problem.**

Transformer-based models such as DistilRoBERTa-base and BETO achieved the strongest overall performance in support detection tasks, particularly when contextual understanding and semantic ambiguity played a central role. However, traditional lexical approaches based on TF-IDF combined with Logistic Regression or SVC remained highly competitive and, in some cases, superior when evaluated using Macro-F1 under class imbalance conditions. This was especially evident in the Spanish target identification task, where explicit lexical and syntactic markers proved more informative than deep contextual embeddings.

The experiments also revealed important limitations associated with topic bias and dataset imbalance. **Although preprocessing strategies such as orthographic normalization and entity masking reduced direct memorization of topical entities, the models remained influenced by the highly centralized geopolitical nature of the dataset.** Furthermore, Transformer-based architectures showed reduced robustness for minority classes despite class-weighting strategies, suggesting that high global accuracy does not necessarily translate into balanced classification performance.

Overall, the findings indicate that social support analysis is a multi-layered problem requiring different modeling strategies depending on the target task. **Traditional sparse representations remain highly effective for structurally explicit classification problems, while contextual Transformers provide stronger semantic generalization for ambiguous or context-dependent support detection.** Future work should explore larger and more diverse datasets, improved imbalance mitigation strategies, and the comparative effectiveness of multilingual versus specialized monolingual architectures in cross-domain social discourse analysis.

Acknowledgments

The authors would like to thank Moein Shahiki Tash for the support and guidance provided throughout the experimentation and writing phases of this work.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) and Gemini in order to: grammar and spelling check, and paraphrase and reword. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

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

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

[4] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[5] 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).

[6] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[7] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[8] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. doi:10.1162/tacl_a_00051.

[10] M. S. Tash, Z. Ahani, L. Ramos, et al., Linguistic analysis in different types of support using liwc for spanish and english online communities, Research Square Preprint (2025). URL: https://doi.org/10.21203/rs.3.rs-9557287/v1. doi:10.21203/rs.3.rs-9557287/v1.

[11] A. Bonet-Jover, J. A. Gonz'alez-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.

[12] 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.

[13] B. L. Monroe, M. P. Colaresi, K. M. Quinn, Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict, Political Analysis 16 (2008) 372–403. doi:10.1093/pan/mpn018.

[14] W. Garbe, 1000x faster spelling correction algorithm, https://github.com/wolfgarbe/SymSpell, 2012. Symmetric Delete spelling correction algorithm; software/blog post, no formal peer-reviewed publication exists.

[15] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. URL: https://doi.org/10.5281/zenodo.1212303. doi:10.5281/zenodo.1212303.