Explainable Social Support Detection in Bilingual YouTube Comments: Cross-Lingual SHAP Analysis of Fine-Tuned XLM-RoBERTa

Melany Jocelyn Monroy Santoyo[1,*]

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

Abstract

Detecting social support in online discourse requires understanding communicative intent rather than surface-level sentiment. We study this problem on YouTube comments in Spanish and English, structured as three hierarchical subtasks: (1) binary supportive vs. non-supportive detection, (2) individual vs. group-targeted support identification, and (3) classification of the specific supported community among six categories (Nation, LGBTQ, Black Community, Women, Religion, Other). We fine-tune XLM-RoBERTa-base with Class-Balanced Focal Loss (β=0.9, γ=2) to handle severe class imbalance (up to 52× between the most and least frequent categories), and compare it against a TF-IDF + logistic regression baseline. We further apply SHAP (SHapley Additive exPlanations) to produce token-level explanations of model predictions across both languages. Our cross-lingual explainability analysis reveals that English supportive language relies on religious solidarity tokens (*muslim*, *islam*, *peace*), while Spanish support is driven by emotional and communal vocabulary (*hermano*, *orgullo*, *Dios*). On a stratified 20% held-out validation split, XLM-RoBERTa achieves macro F1 scores of 0.8255, 0.8226, and 0.8797 for Tasks 1–3, outperforming the baseline on all three subtasks. On the official Codabench test set, our combined submission obtained an aggregate macro F1 of 0.8032 (English) and 0.8541 (Spanish).

Keywords

social support detection, explainability, SHAP, cross-lingual analysis, multilingual NLP, class imbalance

1. Introduction

Social support—the provision of encouragement, care, admiration, help, or solidarity toward someone—is a fundamental aspect of human communication with well-documented benefits for mental health and well-being [1]. In the digital age, social media platforms have become primary venues where individuals express and receive such support. YouTube, as one of the largest user-generated content platforms, hosts millions of comments where users show solidarity with individuals and communities. Detecting supportive language in these comments has practical applications in community moderation, mental health monitoring, and social science research.

From a Natural Language Processing (NLP) perspective, social support detection is distinct from 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 [2]. Consider these simple examples:

- **Supportive (individual):** *"You are so brave, keep fighting!"* — encouragement directed at a specific person.
- **Supportive (group):** *"Proud of all the women making history today"* — solidarity directed at a community (Women).
- **Non-supportive:** *"This video is boring"* — an opinion with no supportive intent.

Detecting this distinction requires models that capture communicative intent beyond surface-level affect. Furthermore, supportive language may manifest differently across languages and cultures: the

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

0249339@up.edu.mx (M. J. M. Santoyo)

lexical and pragmatic patterns used to express solidarity in English may differ substantially from those in Spanish.

In this work, we address bilingual social support detection in YouTube comments in English and Spanish, as part of the Social Support Detection (SSD) shared task at IberLEF 2026 [3, 4]. The task builds on a line of work introducing and extending annotated datasets for this problem [7, 5, 6, 8]. The task is structured as three hierarchical subtasks of increasing granularity:

1. **Support Detection:** Binary classification of comments as supportive or non-supportive.
2. **Target Identification:** For supportive comments, distinguishing whether the support targets a specific individual or a collective group.
3. **Group Target Classification:** For group-targeted comments, identifying the specific community being supported among six categories: Nation, LGBTQ, Black Community, Women, Religion, and Other.

While prior work on this task has focused primarily on maximizing classification performance, a critical gap remains: *understanding what the models learn.* High accuracy is valuable, but without interpretability, practitioners cannot verify whether models rely on meaningful linguistic cues or superficial artifacts, nor can social scientists derive actionable insights about how support is expressed across languages and communities.

This leads to our central research question:

RQ: *Do the linguistic patterns that drive social support detection differ systematically between English and Spanish, and across supported community categories?*

To address this question, we fine-tune XLM-RoBERTa-base [9] on the bilingual dataset, benchmark it against a classical TF-IDF baseline, and apply SHAP (SHapley Additive exPlanations) [10] to produce token-level explanations of model predictions. Our explainability analysis compares the most influential tokens for each class and language, revealing cross-lingual patterns in how support is expressed toward different communities.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 characterizes the dataset. Section 4 describes our methodology. Section 5 presents experimental results. Section 6 provides the cross-lingual explainability analysis. Section 7 discusses findings and limitations. Section 8 concludes.

2. Related Work

2.1. Social Support Detection

Social support detection emerged from the intersection of computational linguistics and health informatics. Cutrona and Suhr [2] established foundational taxonomies of social support types (emotional, informational, esteem, network, and tangible), which have informed subsequent NLP formulations.

Tash et al. [7] introduced a benchmark for social support detection in Spanish social media, comprising Spanish YouTube comments annotated across three hierarchical subtasks (binary support, target type, group category). They evaluated traditional, deep learning, and transformer baselines, including BETO [12] and multilingual BERT [11]. Ahani et al. [5] formalized Social Support Detection as an NLP task and extended the setting to English, evaluating feature combinations spanning linguistic, psycholinguistic, emotional, and sentiment information alongside neural models. Kolesnikova et al. [6] evaluated advanced techniques on the bilingual dataset, including transformer models and zero-shot approaches with GPT-3, GPT-4, and GPT-4o, reporting that transformer-based methods with class balancing outperformed traditional baselines on fine-grained group classification.

2.2. Explainability in NLP

Explainable AI (XAI) methods are increasingly important for validating NLP model decisions. SHAP [10], based on Shapley values from cooperative game theory, provides theoretically grounded feature attributions satisfying local accuracy, missingness, and consistency. Ribeiro et al. [16] introduced LIME, which approximates model behavior locally with interpretable surrogates; while efficient, it offers weaker theoretical guarantees than SHAP [17]. SHAP has been applied to sentiment analysis [18] and hate speech detection [19]. Kokalj et al. [18] found that SHAP attributions aligned better with human judgments than attention weights, while Mathew et al. [19] used SHAP to reveal that hate speech classifiers often relied on identity terms rather than genuinely hateful language.

To our knowledge, no prior work has applied SHAP-based cross-lingual explainability analysis to social support detection, making this a novel contribution of our work.

2.3. Multilingual Transformers and Class Imbalance

XLM-RoBERTa [9] was pre-trained on 2.5TB of CommonCrawl data covering 100 languages, demonstrating strong cross-lingual transfer and outperforming multilingual BERT on benchmarks such as XNLI. For Spanish, BETO [12] has established strong baselines, though multilingual models often benefit from cross-lingual signal when data is sparse in one language [13]. To handle imbalance, Class-Balanced Focal Loss [14] combines the effective-number weighting of class-balanced loss with the hard-example focusing of Focal Loss [15]: the effective number of samples $E(n) = (1 - \beta^n)/(1 - \beta)$ provides a better weighting scheme than simple inverse frequency. We apply this loss to Task 3, where imbalance ratios reach 52× between the largest and smallest classes.

3. Dataset

The dataset consists of 10,598 YouTube comments distributed across English (7,998) and Spanish (2,600), collected from videos related to social and political topics. Comments are annotated across three hierarchical subtasks, where each subsequent task operates on a subset of the previous one.

3.1. Dataset Statistics

Table 1 summarizes the dataset size by task and language. The hierarchical structure progressively filters the data: Task 2 includes only supportive comments (2,385 of 10,598), while Task 3 includes only group-targeted supportive comments (1,906 of 2,385).

Table 1
Dataset size by task and language.

Task (Filter)	English	Spanish	Total
Task 1 (all data)	7,998	2,600	10,598
Task 2 (supportive only)	1,795	590	2,385
Task 3 (group only)	1,450	456	1,906

Table 2 shows the detailed class distribution. Severe imbalance is evident in Task 3: in English, Nation (786 samples) has 52× more examples than Religion (15 samples). The Spanish subset exhibits a different distribution, with LGBTQ as the most frequent category (196 samples) rather than Nation.

3.2. Text Characteristics

Comments exhibit typical social media characteristics: informal register, slang, and platform-specific conventions. English comments average 19 words (median 13, max 157), while Spanish comments average 25 words (median 13, max 744). Approximately 65% of Spanish comments contain non-ASCII

Table 2
Detailed class distributions for all three tasks. Bold indicates the most frequent class; italic indicates the least frequent class per language.

Task	Lang	Class	Count	%
Task 1	EN	Non-Supportive	6,203	77.6%
	EN	Supportive	1,795	22.4%
	ES	Non-Supportive	2,010	77.3%
	ES	Supportive	590	22.7%
Task 2	EN	Individual	345	19.2%
	EN	Group	1,450	80.8%
	ES	Individual	134	22.7%
	ES	Group	456	77.3%
Task 3	EN	**Nation**	**786**	54.2%
	EN	Other	416	28.7%
	EN	LGBTQ	123	8.5%
	EN	Black Community	91	6.3%
	EN	*Women*	*19*	1.3%
	EN	*Religion*	*15*	1.0%
	ES	**LGBTQ**	**196**	43.0%
	ES	Other	81	17.8%
	ES	Religion	60	13.2%
	ES	Women	48	10.5%
	ES	Nation	47	10.3%
	ES	*Black Community*	*24*	5.3%

characters (accented letters, emoji), compared to less than 1% of English comments, reflecting the preprocessed nature of the English data where emojis were converted to text descriptions. Figure 1 shows the Task 3 category distribution and imbalance ratios by language.

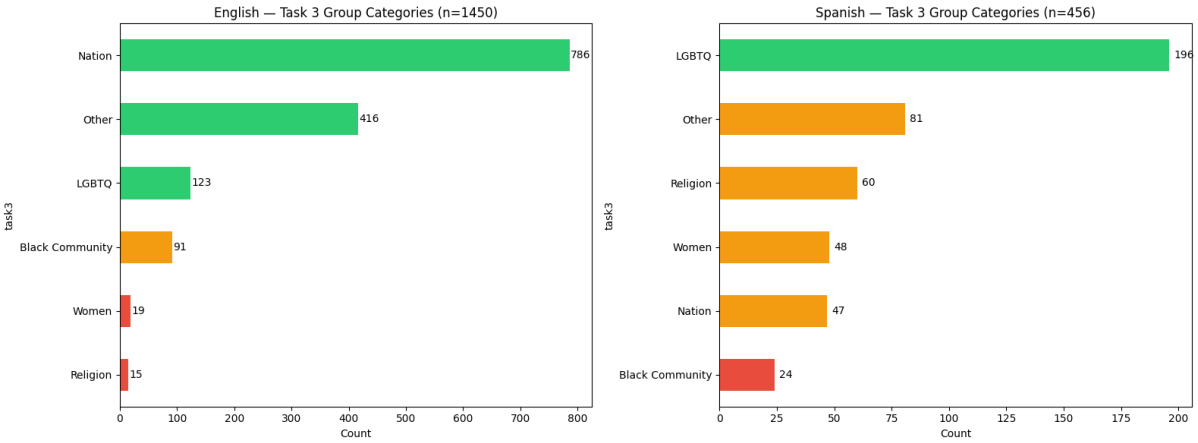

Figure 1: Task 3 group category distribution by language. English is dominated by Nation (786), while Spanish is dominated by LGBTQ (196). Red indicates fewer than 30 samples, orange fewer than 100, green above 100.

3.3. Annotation Scheme

The annotation follows a hierarchical structure. A comment is first labeled as supportive or non-supportive (Task 1); supportive comments that explicitly promote violence or harm are labeled as non-supportive. For supportive comments, annotators determine whether the support targets a specific individual or a collective group (Task 2). For group-targeted comments, the specific community is identified among the six Task 3 categories.

4. Methodology

4.1. Data Preprocessing

Labels are normalized across languages (English uses "Supportive"/"Non-Supportive" while Spanish uses "Support"/"Non Support") and the datasets are merged with a language tag. We perform a stratified 80/20 train/validation split, stratifying jointly by label and language to ensure proportional representation. This yields training/validation splits of 8,478/2,120 for Task 1, 1,908/477 for Task 2, and 1,524/382 for Task 3. The same split is used for the baseline and the transformer to ensure a fair comparison.

4.2. Baseline: TF-IDF + Logistic Regression

As a classical baseline, we train a TF-IDF vectorizer (word unigrams and bigrams, sublinear term frequency, minimum document frequency of 2, up to 30,000 features) followed by a logistic regression classifier with balanced class weights. This provides a strong, interpretable, non-neural reference point that requires no pre-training, isolating the contribution of contextual representation learning in the transformer.

4.3. Model: XLM-RoBERTa

We fine-tune XLM-RoBERTa-base [9] (279M parameters), a multilingual transformer pre-trained on 2.5TB of CommonCrawl data spanning 100 languages. We train three separate classification heads, one per subtask: a 2-class head for Task 1 (all 10,598 comments), a 2-class head for Task 2 (2,385 supportive comments), and a 6-class head for Task 3 (1,906 group-targeted comments).

Training configuration: maximum sequence length 128, batch size 32, learning rate 2×10^{-5} with linear warmup (10%) and linear decay, AdamW optimizer with weight decay 0.01, gradient clipping at 1.0, early stopping with patience 3 (Tasks 1–2) and 5 (Task 3), maximum 10–15 epochs. All experiments use random seed 42.

4.4. Loss Functions

For Tasks 1 and 2, which have moderate class imbalance ($\sim$77/23 split), we use weighted cross-entropy with inverse-frequency class weights. For Task 3, where imbalance is extreme (52$\times$ between Nation and Religion), we employ Class-Balanced Focal Loss [14]. The class-balanced weight is $w_y = 1/E(n_y)$ with $E(n_y) = (1 - \beta^{n_y})/(1 - \beta)$, normalized across classes; combined with the focal modulation $(1 - p_t)^\gamma$ with $\gamma = 2$, this yields $\mathcal{L}_{\text{CB-Focal}}(p_t) = -w_y(1 - p_t)^\gamma \log(p_t)$. We set $\beta = 0.9$ rather than the higher values (e.g., 0.9999) often suggested in the original formulation, as smaller class sizes in our setting make a less aggressive re-weighting more stable in practice.

4.5. SHAP Explainability

After training, we apply SHAP [10] with the partition masker for text to compute token-level Shapley values. We choose SHAP over alternatives such as LIME [16] or Integrated Gradients because SHAP provides theoretically grounded attributions with local accuracy and consistency guarantees, its partition masker is well suited to subword-tokenized text, and it has become a standard, comparable method for transformer interpretability [18, 17]. For each correctly classified validation example, SHAP assigns each token a value indicating its contribution to the predicted class probability. We sample balanced subsets of correctly classified examples (20 per language-class combination for Tasks 1–2; 10 for Task 3) so that explanations reflect genuine model behavior rather than error patterns. From these we compute (i) global token importance (mean absolute SHAP value per token), (ii) a cross-lingual comparison aggregated separately by language, and (iii) per-category token profiles for Task 3.

5. Results

5.1. Baseline vs. Transformer

Table 3 presents validation macro F1 for both the TF-IDF baseline and XLM-RoBERTa on the same held-out split. XLM-RoBERTa outperforms the baseline on all three subtasks, with the largest margin on Task 3 (+0.126), the most fine-grained and imbalanced subtask. This confirms that contextual representation learning provides its greatest benefit precisely where lexical features are least sufficient, while the competitive baseline scores (0.75–0.81) indicate that the tasks contain strong surface-level signal as well.

Table 3
Validation macro F1: TF-IDF + logistic regression baseline vs. XLM-RoBERTa, on the same stratified split.

Task	Baseline	XLM-R	Δ
Task 1 (Support Detection)	0.7874	0.8255	+0.038
Task 2 (Target Identification)	0.8099	0.8226	+0.013
Task 3 (Group Classification)	0.7536	0.8797	+0.126

Table 4 reports the XLM-RoBERTa scores broken down by language.

Table 4
XLM-RoBERTa validation macro F1 by task and language.

Task	Overall	English	Spanish
Task 1	0.8255	0.8210	0.8378
Task 2	0.8226	0.8385	0.7780
Task 3	0.8797	0.7707	0.8886

5.2. Official Test Set Results

We submitted predictions for all three subtasks to the official Codabench evaluation platform as a single combined submission per language. The platform reported an aggregate macro F1 of **0.8032** for English and **0.8541** for Spanish. As the leaderboard returns a single aggregate score per language rather than per-subtask breakdowns, the per-task and per-class analyses below are reported on our stratified 20% held-out validation split, for which we have full label access.

5.3. Per-Class Performance and Statistical Note

Table 5 shows per-class F1 for Task 3. The English Religion score (0.00 F1) and the perfect English Women score (1.00 F1) should be read with caution, as these classes have only 3 and 4 validation samples respectively. We do not report formal significance tests: with as few as 3–4 validation instances in the smallest categories, a single prediction changes macro F1 substantially, so bootstrap or permutation intervals would be too wide to support meaningful pairwise claims. We therefore treat small-class results as indicative and rely on the consistent cross-task margin over the baseline (Table 3) as our main quantitative evidence.

Figure 2 shows the validation confusion matrices for the three subtasks. The dominant Task 3 error is the confusion between Nation and Other, examined qualitatively in Section 7.

All three tasks converge within 7–12 epochs (Task 1 at epoch 10, Task 2 at epoch 3 with early stopping at 6, Task 3 at epoch 7 with early stopping at 12).

Table 5
Task 3 per-class F1 scores by language (validation split).

Category	EN F1	ES F1
Black Community	0.9444	0.8000
LGBTQ	0.9600	0.9114
Nation	0.8911	0.9000
Other	0.8287	0.9032
Religion	0.0000	0.8696
Women	1.0000	0.9474
Macro F1	0.7707	0.8886

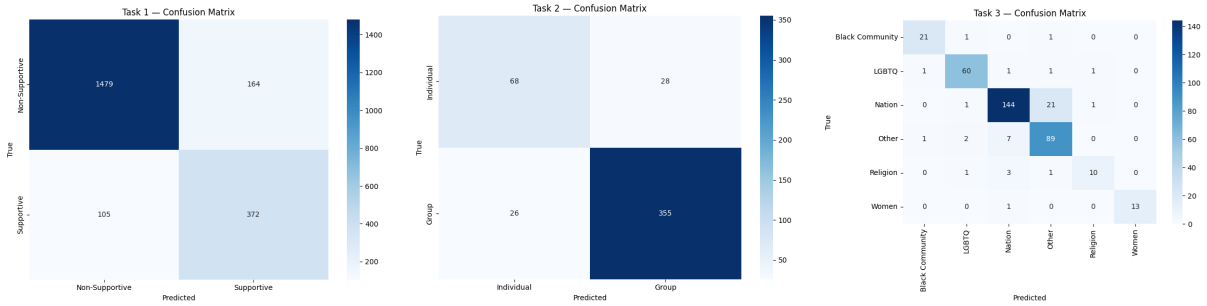

Figure 2: Confusion matrices on the validation set for Task 1 (left), Task 2 (center), and Task 3 (right). The main Task 3 confusion is between Nation and Other.

6. Cross-Lingual Explainability Analysis

This section presents the main contribution of our work: a systematic comparison of the linguistic patterns that drive model predictions across English and Spanish.

6.1. Task 1: What Makes a Comment Supportive?

Figure 3 shows the global top tokens driving the "Supportive" classification. The cross-lingual comparison reveals markedly different patterns. In English, the model relies on tokens related to religious solidarity and social movements: *muslim* (0.075), *alla* (0.048), *peace* (0.043), *islami* (0.040), and *religion* (0.024) dominate, reflecting the English data's focus on interfaith solidarity and social justice discourse. In Spanish, tokens such as *hermano* (brother), *orgullo* (pride), *Dios* (God), *México*, and *Fe* (faith) appear prominently, suggesting that Spanish supportive language is expressed through familial, national, and religious registers rather than the activism-oriented framing observed in English.

6.2. Task 3: Category-Specific Patterns

Figure 4 shows the most important tokens for each of the six group target categories. The per-class analysis reveals category-specific vocabulary that aligns with expected thematic content: Black Community is dominated by *negro* (0.50), *black* (0.19), and *rac* (0.16); LGBTQ by *homosexual* (0.52), *homo* (0.48), and *gay* (0.43); Nation by *SPORT* (0.42), *MEX* (0.33), and *palestin* (0.26); Religion by *Jesus* (0.43), *Cristo* (0.43), and *religio* (0.31); and Women by *mujeres* (0.51), *moms* (0.41), and *chica* (0.34). The Other category is heterogeneous, mixing *mujeres*, *gay*, and *SPORT*. These results indicate that the model captures meaningful linguistic distinctions rather than superficial artifacts.

6.3. Cross-Lingual Comparison and the Effect of Imbalance

The cross-lingual comparison reveals an important asymmetry tied to data availability. For English Religion, SHAP values are near zero: with only 3 validation samples and 0.00 F1, the model has

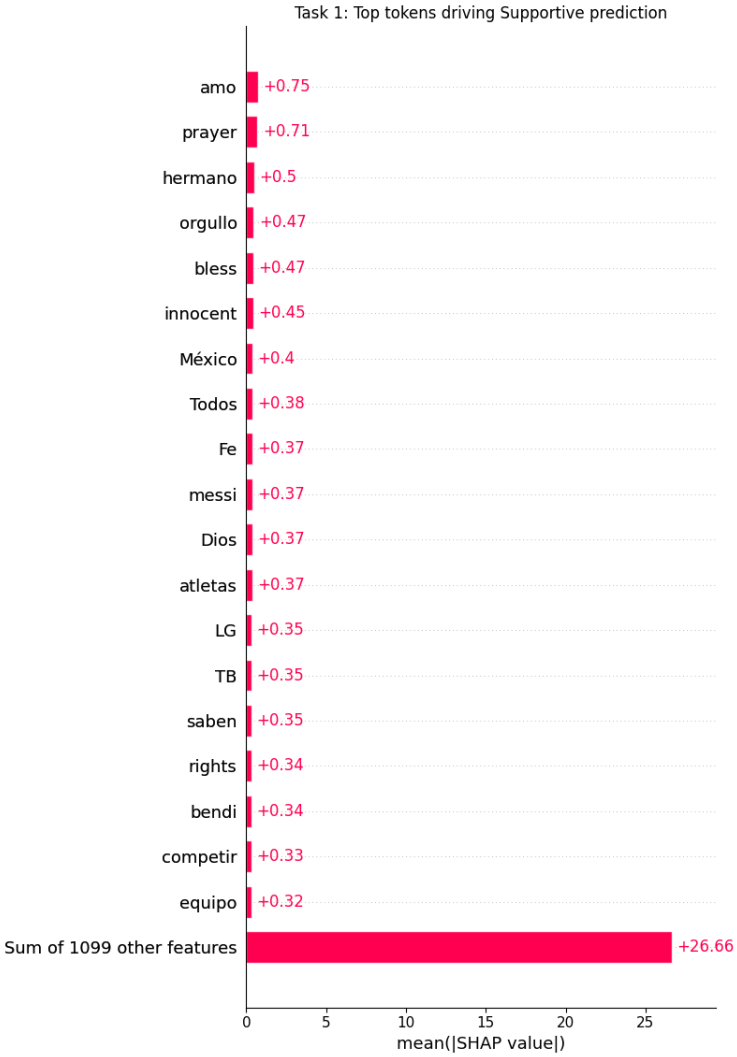

Figure 3: Top 20 tokens by mean absolute SHAP value for the Supportive class (both languages combined). Spanish tokens (*amo*, *hermano*, *orgullo*, *México*, *Dios*) and English tokens (*prayer*, *bless*, *innocent*, *rights*) both appear.

no interpretable signal for this class. In contrast, Spanish Religion (12 validation samples, 0.87 F1) shows clear signal with tokens like *innocent*. This demonstrates that class imbalance affects not only performance but also interpretability. For Nation, English relies on subword fragments of sports and geopolitical terms, while Spanish shows a broader, lower-magnitude token distribution, suggesting that Spanish national support is expressed through more diverse vocabulary. Table 6 consolidates the dominant token themes by language and task.

Table 6
Summary of dominant token themes by language and task.

Category	English tokens	Spanish tokens
Supportive (T1)	muslim, peace, islam	hermano, orgullo, Dios
Black Community (T3)	negro, black, rac	Comunidad, negra
LGBTQ (T3)	homosexual, homo, gay	homosexual, sex, amor
Nation (T3)	SPORT, MEX, palestin	México, atletas, equipo
Religion (T3)	(insufficient data)	Jesus, Cristo, religio
Women (T3)	mujeres, children	mujeres, chica, fem

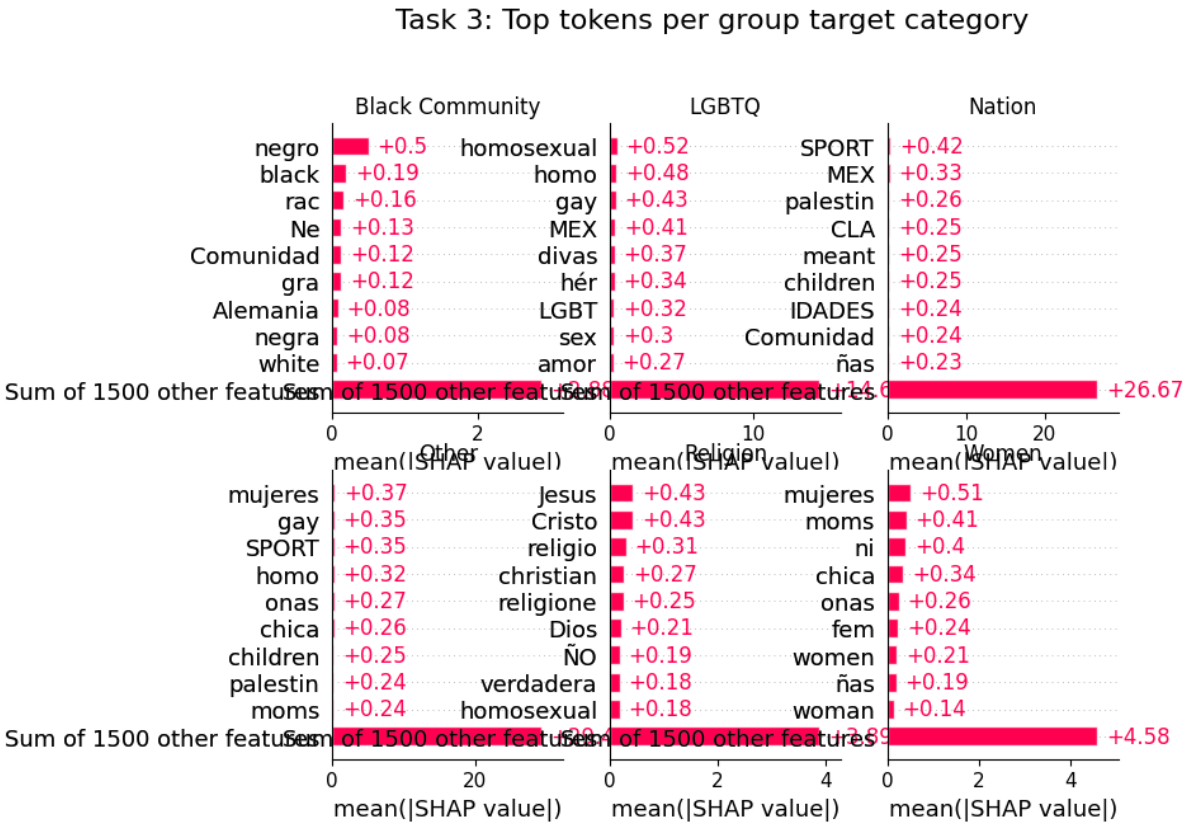

Figure 4: Top 10 tokens by mean absolute SHAP value for each of the six group categories in Task 3.

7. Discussion

Cross-lingual patterns. Supportive language is not expressed through direct translation equivalents across languages. English support is associated with religious-solidarity discourse (*muslim*, *islam*, *peace*), likely reflecting interfaith solidarity content in the sampled comments, whereas Spanish support is expressed through familial and communal vocabulary (*hermano*, *orgullo*) alongside religious terms (*Dios*, *Cristo*). The Nation category diverges most: English links to sports and Palestinian solidarity (*SPORT*, *palestin*), while Spanish centers on Mexican national identity (*MEX*, *México*). Multilingual support-detection models should therefore not assume cross-lingual feature equivalence.

Qualitative error analysis. The dominant Task 3 error is the confusion between Nation and Other (44 of the misclassified validation cases). Inspecting these reveals a genuine annotation ambiguity rather than a modelling failure: comments about Gaza and Palestine—e.g., *"makes [me] feel sad to see children … died [in] gaza … support gaza palestine"*—are gold-labelled inconsistently between Other and Nation, and the model's predictions often match the alternative reading. Such borderline cases, where support for a people and support for a nation overlap, place an upper bound on achievable Task 3 accuracy and complement the SHAP finding that Nation relies heavily on geopolitical tokens.

Validation vs. test and the role of CB-Focal. The gap between validation (0.8255 on Task 1) and the aggregate Codabench scores (0.8032 EN / 0.8541 ES) is expected: the official metric folds all three hierarchical subtasks into one per-language number, so errors propagate from Task 1 downward. CB-Focal Loss with β=0.9 was important for Task 3: without balancing, the model collapses toward the majority class, whereas with it several minority classes are learned. Still, balancing cannot overcome extreme scarcity—English Religion (15 total examples) remains at 0.00 F1—so the smallest-class scores should be read as indicative.

Limitations. Our study has four main limitations. First, we evaluate a single transformer (XLM-RoBERTa-base) against one classical baseline; monolingual models such as BETO are not compared, so conclusions about architecture are bounded. Second, the SHAP analysis uses correctly classified examples and operates on SentencePiece subword tokens, which fragment words differently across languages. Third, several Task 3 categories have very few validation samples, limiting the reliability of per-class scores. Finally, the data is drawn from YouTube comments on specific social and political topics, which may limit generalizability to other platforms or domains.

8. Conclusion

We presented an explainable approach to social support detection in bilingual YouTube comments. Fine-tuned XLM-RoBERTa outperforms a TF-IDF baseline on all three hierarchical subtasks—most clearly on the imbalanced group-classification task—and obtains aggregate test macro F1 of 0.8032 (English) and 0.8541 (Spanish) on the official Codabench leaderboard. The central contribution is a cross-lingual SHAP analysis showing that (i) English supportive language relies on religious-solidarity vocabulary while Spanish favors familial, communal, and national expressions; (ii) per-category tokens align with expected thematic content; and (iii) class imbalance degrades not only performance but also interpretability, visible as near-zero SHAP signal for the data-starved English Religion class. Future work should compare monolingual architectures, extend the analysis to more languages, and test whether SHAP-based insights can guide data augmentation.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools to assist with language editing (grammar and clarity) and to help debug and refactor portions of the experimental code. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] J. S. House, *Work Stress and Social Support*, Addison-Wesley, Reading, MA, 1981.

[2] C. E. Cutrona, J. A. Suhr, Controllability of stressful events and satisfaction with spouse support behaviors, *Communication Research* 19 (2) (1992) 154–174.

[3] M. Shahiki Tash, L. Ramos, Z. Ahani, A. Y. Barrera Animas, F. Balouchzahi, O. Kolesnikova, G. Sidorov, A. Gelbukh, R. Monroy, F. Vargas, Overview of SSD-2026 at IberLEF 2026: Social Support Detection, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026)*, CEUR-WS.org, 2026.

[4] 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.

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

[6] 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 (10) (2025) e43437. doi:10.1016/j.heliyon.2025.e43437.

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

[8] 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 preprint (2024). doi:10.21203/rs.3.rs-9557287/v1.

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

[10] S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

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

[13] A. Ebrahimi, J. Lee, K. Lee, et al., On the use of multilingual versus monolingual models in NLP tasks, *Computational Linguistics* 47 (2021) 829–858.

[14] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: *Proceedings of CVPR*, 2019, pp. 9268–9277.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proceedings of ICCV*, 2017, pp. 2980–2988.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: *Proceedings of KDD*, ACM, 2016, pp. 1135–1144.

[17] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2020. christophm.github.io/interpretable-ml-book.

[18] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, M. Robnik-Šikonja, BERT meets Shapley: Extending SHAP explanations to transformer-based classifiers, in: *Proceedings of the EACL Student Research Workshop*, 2021, pp. 16–21.

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A benchmark dataset for explainable hate speech detection, in: *Proceedings of AAAI*, 2021, pp. 14867–14875.