Leveraging Multilingual Transformers for Social Support Detection and Community Identification in Social Media

Joel Vázquez Anaya[1,*]

[1]Universidad Panamericana, Ciudad de México, México

Abstract

This paper presents a multilingual transformer-based system submitted to the IberLEF 2026 Social Support Detection (SSD-2026) shared task, which frames prosocial language understanding as a three-level hierarchical classification problem over English and Spanish social media texts. The proposed system uses XLM-RoBERTa-base in a zero-shot classification setting, relying on natural-language hypothesis templates to transfer multilingual knowledge from pretraining without task-specific fine-tuning. The pipeline addresses: (i) binary detection of supportive versus non-supportive messages, (ii) identification of individual- versus group-directed support, and (iii) categorization of the specific community supported among six classes: Nation, Other, LGBTQ, Black Community, Religion, and Women. The system was applied to the official SSD-2026 test sets, containing 2,000 English and 651 Spanish comments. The submitted system was evaluated through the SSD-2026 official leaderboard and achieved an average macro-F1 score of 0.8151 across the six language-subtask evaluations. In addition to these official macro-F1 scores, this paper analyzes the prediction distributions produced by the zero-shot pipeline. These distributions reveal important limitations, including overprediction of support in English and underprediction of several minority communities, motivating supervised fine-tuning with the SSD-2026 training data in future work.

Keywords

social support detection, SSD-2026, IberLEF 2026, XLM-RoBERTa, zero-shot classification, multilingual NLP, community identification

1. Introduction

The rapid expansion of social media platforms has transformed how individuals seek and provide emotional assistance online. While a substantial body of NLP research has focused on detecting harmful content, such as hate speech, cyberbullying, and offensive language, there is also growing recognition of the importance of identifying and promoting prosocial behavior as a complementary research direction. Among the forms of prosocial interaction observable in online discourse, social support stands out as a critical element for psychological resilience, offering digital assistance to vulnerable individuals and marginalized communities.

Detecting social support is non-trivial: it requires distinguishing between the active expression of empathy, solidarity, encouragement, or instrumental aid and mere positive sentiment, while also contending with multilingual contexts, informal registers, emojis, and cultural slang. The SSD-2026 shared task at IberLEF 2026 formalizes this challenge as a three-level hierarchical classification problem over English and Spanish social media comments [1, 2]. The dataset and annotation framework are grounded in prior work on English social support detection [3], Spanish social support detection [5], transformer-based social support detection [4], and LIWC-based linguistic analysis of support types across languages [6].

The shared task is organized as follows:

1. **Subtask 1:** binary detection of supportive versus non-supportive messages.
2. **Subtask 2:** given a supportive message, determining whether support targets an individual or a group.
3. **Subtask 3:** identifying the specific community receiving support among Nation, Other, LGBTQ, Black Community, Religion, and Women.

IberLEF 2026, León, Spain

*Corresponding author.

0201031@up.edu.mx (J. V. Anaya)

This paper describes a zero-shot baseline system based on XLM-RoBERTa-base. The choice of a zero-shot setup was motivated by the goal of evaluating whether a multilingual pretrained model could produce a complete valid submission without task-specific supervised fine-tuning. This version does not claim that training data were unavailable to all participants; rather, it documents a deliberate zero-shot configuration and its limitations.

2. Related Work

Social support detection has recently emerged as a distinct research direction at the intersection of sentiment analysis, affective computing, and computational social science. Ahani et al. [3] introduced social support detection from social media texts as a structured NLP problem and provided a foundation for the annotation scheme used in the shared task. Kolesnikova et al. [4] compared advanced machine learning and transformer-based methods for social support detection, showing that contextual language models can outperform traditional baselines. Tash et al. [5] extended the task to Spanish social media text, while Shahiki-Tash et al. [6] analyzed linguistic and psycholinguistic markers of support types in Spanish and English online communities.

The SSD-2026 shared task [1] builds directly on these resources and evaluates systems in a bilingual setting at IberLEF 2026 [2]. The task is particularly challenging because it combines binary support detection, target identification, and fine-grained community classification in a hierarchical structure.

Zero-shot classification via natural language inference (NLI) has become a practical method for applying pretrained models to new classification tasks without labeled data. In this paradigm, the input text is paired with a natural-language hypothesis describing a class, and the model estimates whether the input entails that hypothesis. This formulation is useful in low-resource settings, but its performance depends strongly on the phrasing of candidate labels and may introduce systematic biases across languages and categories.

3. Methodology

3.1. Model

The proposed system uses XLM-RoBERTa-base, a multilingual transformer model pretrained on large multilingual corpora [7]. Its shared cross-lingual representations make it suitable for a bilingual task spanning English and Spanish social media data.

Classification was performed using a zero-shot classification pipeline. For each input comment, candidate labels were expressed as natural-language hypotheses. The model selected the class with the highest entailment score. No task-specific supervised fine-tuning was performed in this version of the system.

3.2. Hypothesis Templates

The following templates were used:

- **Subtask 1:** “supportive message” vs. “non-supportive message”.
- **Subtask 2:** “support for an individual person” vs. “support for a group or community”.
- **Subtask 3:** “support for Black community”, “support for LGBTQ community”, “support for a nation or country”, “other type of support”, “support for a religious group”, and “support for women”.

3.3. Cascaded Inference

The three subtasks were addressed through a sequential cascade. Subtask 1 was applied to all comments. Subtask 2 was applied only to comments predicted as supportive. Subtask 3 was applied only to

comments predicted as supportive and group-directed. Comments predicted as non-supportive received neutral downstream fallback values in order to satisfy the required output format. This cascade respects the semantic structure of the task, but it also propagates errors from earlier decisions into downstream predictions.

3.4. Implementation

The pipeline was implemented in Python using the Hugging Face `transformers` library. Inference was performed in batches of 64 on a single NVIDIA T4 GPU with 16 GB of VRAM.

4. Experimental Setup

The system was applied to the official SSD-2026 test sets: 2,000 English comments and 651 Spanish comments. Official leaderboard macro-F1 scores for the submitted system are reported in Table 1. Prediction distributions are additionally reported in Tables 2–4 as descriptive output analysis. The official scores and prediction counts measure different aspects of the system: the former evaluate performance against gold labels, while the latter describe the output behavior of the zero-shot pipeline.

5. Results

Table 1 reports the official macro-F1 scores for the submitted system across the six SSD-2026 language-subtask evaluations. The system achieved an overall average macro-F1 of 0.8151. Spanish performance was slightly stronger on average (0.8261) than English performance (0.8040). These official scores are reported without leaderboard ranks; the following tables focus on the prediction distributions, which help explain the zero-shot behavior behind the submission.

Table 1
Official SSD-2026 macro-F1 scores for the submitted system.

Language	Subtask	Macro-F1
Spanish	Subtask 1	0.8514
Spanish	Subtask 2	0.8123
Spanish	Subtask 3	0.8148
English	Subtask 1	0.7798
English	Subtask 2	0.8225
English	Subtask 3	0.8096
Average		0.8151

Tables 2–4 report the prediction distributions produced by the zero-shot system.

Table 2
Subtask 1 – Support Detection: prediction counts and proportions.

Language	Total	Support	Non-Support	% Support
English	2,000	1,976	24	98.8
Spanish	651	509	142	78.2

The most salient observation is the overprediction of support in English: 98.8% of English comments were classified as supportive. This is unlikely to reflect the true distribution of social support in a general social media corpus and suggests that the English hypothesis pair biased the model toward the positive-support class. Spanish predictions were more conservative, with 78.2% classified as supportive, but this rate is still high for a social support detection dataset.

Table 3
Subtask 2 – Target Identification: distribution among comments predicted as supportive.

Language	Supportive	Individual	Group	% Individual
English	1,976	213	1,763	10.8
Spanish	509	368	141	72.3

For Subtask 2, English comments were mostly classified as group-directed, while Spanish comments were mostly classified as individual-directed. This cross-lingual asymmetry likely reflects template sensitivity rather than a reliable linguistic pattern. For Subtask 3, the Other category dominated in both languages, and the model almost never predicted LGBTQ or Black Community. This underprediction is an important fairness concern because the system may fail to recognize support directed at marginalized communities.

Table 4
Subtask 3 – Community Classification: distribution among comments predicted as group-directed.

Lang.	n	Black	LGBTQ	Nation	Other	Religion	Women
English	1,763	4	0	220	1,073	432	34
Spanish	141	0	0	49	85	5	2

6. Discussion

The official scores and output distributions show that a zero-shot multilingual pipeline can generate a complete valid submission, but they also reveal systematic weaknesses. First, zero-shot templates strongly affect output distributions. Second, the cascade amplifies early errors: if Subtask 1 overpredicts support, Subtasks 2 and 3 are applied to many comments for which downstream labels may not be meaningful. Third, the system underpredicts several minority communities, showing that zero-shot inference is not sufficient for reliable community-level social support detection.

These findings suggest that supervised fine-tuning with SSD-2026 training data is necessary for robust performance. Future versions should train task-specific classification heads for the three subtasks, use class-weighted loss or balanced sampling to address minority classes, and optimize hypothesis templates in both English and Spanish if zero-shot or prompt-based methods are retained as baselines.

7. Limitations

The main limitation of this work is the absence of supervised task-specific fine-tuning. The system therefore cannot learn SSD-2026 decision boundaries directly. A second limitation is template sensitivity: small changes in candidate label wording may produce very different predictions. Third, the system processes each comment in isolation, without thread or conversational context. Finally, although official macro-F1 scores are reported, the prediction-distribution analysis remains limited because it does not provide class-level gold-label error patterns; therefore, the paper should not overinterpret the distributional results as a substitute for detailed per-class evaluation.

8. Conclusion

This paper presented a zero-shot multilingual pipeline for the SSD-2026 shared task at IberLEF 2026. The system uses XLM-RoBERTa-base with natural-language hypothesis templates to address support detection, target identification, and community classification in English and Spanish. The official results show an average macro-F1 of 0.8151 across the six language-subtask evaluations. At the same

time, the distributional analysis shows that zero-shot inference can produce biased output patterns, including support overprediction and weak coverage of minority communities. These limitations motivate supervised fine-tuning, multilingual template optimization, and class-imbalance handling in future work.

Acknowledgments

The author thanks the SSD-2026 and IberLEF 2026 organizers for providing the shared-task dataset, task definition, and evaluation framework.

Declaration on Generative AI

During the preparation of this work, the author used a generative AI (Chat GPT 5.5) assistant to support grammar refinement, academic tone, and formatting of mathematical and LaTeX notation. After using these tools, the author reviewed and edited all content as needed and takes full responsibility for the scientific content and conclusions of this publication.

References

[1] M. Shahiki Tash, L. Ramos, Z. Ahani, A. Y. Barrera-Animas, F. Balouchzahi, O. Kolesnikova, G. Sidorov, A. Gelbukh, R. Monroy, and 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.

[2] A. Bonet-Jover, J. A. González-Barba, and 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.

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

[4] O. Kolesnikova, M. S. Tash, Z. Ahani, A. Agrawal, R. Monroy, and G. Sidorov. Advanced machine learning techniques for social support detection on social media. *Heliyon*, 11(10), 2025. https://doi.org/10.1016/j.heliyon.2025.e43437.

[5] M. S. Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, and G. Sidorov. Online social support detection in Spanish social media texts. *arXiv preprint arXiv:2502.09640*, 2025.

[6] M. Shahiki-Tash, Z. Ahani, L. Ramos, O. Kolesnikova, et al. Linguistic analysis in different types of support using LIWC for Spanish and English online communities. *Research Square*, 2026. https://doi.org/10.21203/rs.3.rs-9557287/v1.

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