Multilingual Transformer Models for Social Support Detection on Social Media Text

Fatima Uroosa[1,†], Muhammad Tayyab Zamir[1,†] and Muhammad Arif[2,*,†]

[1]*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Av. Juan de Dios Bátiz, s/n, Mexico City 07320, Mexico*

[2]*Yale University, 464 Congress Ave 260, New Haven, CT 06519, USA*

Abstract

Social support detection in online communities has become an important research area due to the increasing use of social media platforms for emotional expression and psychological assistance. This study presents a transformer-based approach for multilingual social support detection using Multilingual BERT (mBERT) as part of the SSD-2026 shared task IberLEF 2026. The proposed system was evaluated on English and Spanish social media datasets across multiple classification tasks for identifying supportive interactions. The model was fine-tuned using task-specific preprocessing, tokenization, and supervised learning strategies. Experimental results demonstrate that mBERT achieved competitive performance across multilingual settings, particularly on English-language tasks, while maintaining reasonable cross-lingual generalization. The findings highlight the effectiveness of transformer-based multilingual representations for detecting supportive communication in online environments and provide insights into challenges associated with class imbalance and multilingual social media text classification.

Keywords

Social Support Detection, Multilingual BERT, Mental Health, Natural Language Processing, Transformer Models, Social Media Analysis

1. Introduction

Social support is a key component in achieving better mental health [1] [2] [3] [4]. Findings from earlier studies demonstrate that supportive relationships can have a "buffering effect", meaning they can protect against the detrimental effects of mental illness. Consequently, social support is seen as an important psychological coping mechanism.

According to [5], social support has significant implications for health services and interventions aimed at improving elements of support systems for people at risk of suicidal ideation. Social media platforms such as YouTube, Instagram, Snapchat, TikTok, and Facebook have become integral to daily life globally, transforming how individuals communicate, share information, and seek emotional interaction. [6]. These platforms enable users to observe, engage with, and respond to others through likes, comments, shares, and messages, generating social rewards that reinforce repeated and prolonged usage [4, 7, 8].

With the rise of social platforms, more people are expressing emotions and seeking support online. Social support among these communities has been reported to be linked with improved self-esteem, quality of life, and better well-being. Individuals seek help, advice, and support online [9]. Recent studies have shown that perceived social support plays a significant role in improving mental health outcomes by reducing stress, anxiety, and depressive symptoms. At the same time, even moderate levels of supportive interaction can positively influence emotional well-being when individuals perceive the support as meaningful and accessible [10, 11] [12].

To address these challenges, the Social Support Detection Shared Task at IberLEF 2026 focuses on the automatic identification of supportive interactions in social media comments, including the detection

IberLEF 2026, September 2026, León, Spain

*Corresponding author

†These authors contributed equally.

furoosa2024@cic.ipn.mx (F. Uroosa); mzamir2023@cic.ipn.mx (M. T. Zamir); arifmuhmmand@gmail.com (M. Arif)

of support type and its target (individual or community). In this work, we propose a transformer-based approach for social support detection using multilingual social media data. Specifically, we employ multilingual models including BERT, RoBERTa, and TF-IDF–based classifiers, to classify supportive content in English and Spanish texts. Our study aims to contribute toward effective NLP systems capable of understanding prosocial communication beyond conventional sentiment analysis for real-world online support applications.

Despite recent progress in mental health-related NLP research, multilingual social support detection remains challenging due to linguistic variability, informal social media language, contextual ambiguity, and class imbalance. Many previous studies have focused primarily on English-language datasets or traditional machine learning approaches, limiting their generalizability across multilingual environments. To address these challenges, this study investigates the effectiveness of transformer-based multilingual representations using mBERT for detecting supportive interactions across English and Spanish social media datasets.

The main contributions of this work are as follows: (i) implementation of a multilingual transformer-based framework for social support detection; (ii) evaluation of mBERT on multilingual shared-task datasets involving English and Spanish social media text; and (iii) analysis of model performance across different classification tasks and linguistic settings.

2. Related Work

With the rapid growth of social media platforms, research on identifying supportive and prosocial communication has gained increasing attention in the natural language processing (NLP) community. Unlike hate speech detection, which has been extensively studied, social support detection remains relatively underexplored, particularly in multilingual and Spanish-language social media contexts. Recent studies have investigated the use of machine learning and deep learning approaches for detecting supportive interactions in online conversations.

For instance, transformer-based and large language model (LLM) approaches have shown promising performance in social media tasks. [13] employed a Zero-Shot and transformer-based methods for social support detection using models such as GPT-3 and GPT-4, demonstrating that transformer architectures can effectively capture supportive and community-oriented language patterns. Additionally, recent studies have also explored the relationship between online social support and user behaviour on social media platforms. [14] analyzed self-disclosure and social support interactions among women experiencing infertility on Reddit using BERT-based text classification methods. Their work combined communication and social support theories with NLP techniques to identify different forms of supportive communication in Reddit discussions. Consequently, [15, 16] proposed text-mining approaches for automatically identifying discussion themes and social support categories in online health forums.

Similarly, recent studies have highlighted the effectiveness of transformer models such as BERT, DistilBERT, and DeBERTa for analyzing mental health and supportive communication in social media texts [17, 18] [19, 20]. Other works have also explored transformer-based frameworks for detecting depression, stress, cyberbullying, and harmful online interactions, further demonstrating the capability of contextual language models in understanding complex social media discourse. Additionally, a study by [21, 22] examined the impact of online social support on mental health using the Reddit dataset. Their study employed propensity score matching and qualitative analysis to examine how different forms of support influence future suicidal ideation. The findings showed that emotional, esteem, and network support were associated with reduced suicidal ideation risk, highlighting the importance of supportive communication patterns in online mental health communities.

[23] developed a lexicon-based tweet detection framework to support culturally sensitive Twitter-based social support interventions for African American dementia caregivers. Their study collected millions of tweets related to African American discourse and applied statistical feature selection with Naive Bayes classification to identify linguistic patterns associated with supportive online communities. The authors highlighted the importance of social media analytics and NLP techniques for designing

culturally aware support systems. They suggested that transformer-based models such as BERT could further improve classification performance. A recent study reported a relationship between perceived social support and health-promoting lifestyles among women participating in a breast cancer early-detection program. The findings revealed a significant positive correlation between perceived social support and healthy lifestyle behaviours, including stress management, interpersonal relationships, and self-actualization. The study further emphasized that family and social support contribute to improved psychological well-being, quality of life, and healthier behavioral outcomes. [24]. Another study employed multilingual transformer-based models for detecting racial hoaxes in Hindi-English code-mixed social media data. The results demonstrated the effectiveness of BERT-based architectures in understanding socially sensitive and emotionally nuanced online communication, which is also relevant to social support detection tasks [25]. This study utilizes the dataset released as part of the Social Support Detection Shared Task at IberLEF 2026. The dataset and task formulation are based on previous work on social support detection in social media, including the Spanish social support corpus introduced by [16] and the large-scale social support detection framework proposed by [26]. These resources provide the foundation for multilingual social support detection research and benchmark evaluation.

Several studies have explored the detection of mental health-related content in social media. For instance, [17] used transformer-based models for depression detection in social media posts. In contrast, Bokolo and Liu [19] conducted a comparative analysis of machine learning and transformer models for depression and suicide detection. Similarly, [27] proposed a hybrid CNN–BiLSTM framework for speech emotion recognition, demonstrating the effectiveness of combining spectral and temporal representations for identifying emotional patterns in speech. Their findings highlight the importance of emotion-aware AI systems for understanding human communication, a finding that is also relevant to socially supportive interactions in online environments.

3. Methods

3.0.1. Dataset

This study uses the dataset from the Social Support Detection Shared Task at IberLEF 2026. The dataset comprises social media comments in two languages, English and Spanish, annotated for three classification subtasks. Subtask 1 is a binary classification task that identifies whether a comment is supportive (label: 1) or non-supportive (label: 0). Subtask 2 is a binary classification task that determines the target of support, distinguishing between group-directed (label: 0) and individually directed (label: 1) interactions. Subtask 3 is a multi-class classification task that identifies the specific community targeted by the support, including Black Community (0), LGBTQ (1), Nation (2), Other (3), Religion (4), and Women (5).

The datasets exhibit class imbalance across several categories, which increases the difficulty of accurately classifying minority classes and motivates the use of macro-averaged evaluation metrics.

3.0.2. Preprocessing

All input texts underwent a consistent preprocessing pipeline before model training. Text was first converted to lowercase, after which URLs were removed using regular expressions. Non-alphabetic characters, including punctuation and digits, were subsequently stripped, and any resulting sequences of excess whitespace were collapsed into single spaces. Duplicate entries and rows with missing values were discarded from the training data. These steps were applied uniformly across both the English and Spanish corpora.

3.0.3. Models

Multilingual BERT (mBERT) For the reported Subtasks 1 and 2 across both languages, we fine-tuned bert-base-multilingual-cased (mBERT), a transformer-based language model pre-trained on Wikipedia text spanning 104 languages. mBERT's shared multilingual representation space makes it well-suited to cross-lingual classification without requiring language-specific architectures. The model was loaded with a sequence classification head corresponding to the number of target classes for each subtask, and

RoBERTa For the English Subtask 2 binary classification, we additionally trained roberta-base, a robustly optimized BERT pre-training approach that removes the next-sentence prediction objective and employs dynamic masking. Unlike the multilingual BERT variant, RoBERTa is an English-only model and was applied exclusively to the English portion of the dataset for this subtask.

Table 1
Performance results for Task 1 on the English and Spanish datasets.

Language	Accuracy	Macro F1	Weighted F1
English	0.8630	0.7974	0.8618
Spanish	**0.8925**	**0.8288**	**0.8854**

Table 2
Performance results for Subtask 2 on the English and Spanish datasets.

Language	Accuracy	Macro F1	Weighted F1
English	**0.8730**	0.6622	**0.8634**
Spanish	0.8602	**0.7240**	0.8520

3.1. Training Configuration

All models were tokenized using their respective fast tokenizers with a maximum sequence length of 256 tokens for mBERT experiments and 128 tokens for RoBERTa. Inputs were padded to the maximum length and truncated where necessary. Training was conducted using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and 3 training epochs. A linear learning rate scheduler with no warm-up steps was applied during RoBERTa training. For mBERT, training was managed via the Hugging Face `TrainingArguments` with per-epoch evaluation and model checkpointing. All experiments were executed on an NVIDIA A100 GPU via Google Colab.

3.2. Evaluation

Model performance was assessed on held-out test sets using standard classification metrics: precision, recall, F1-score, and overall accuracy. Both macro-averaged and weighted-averaged metrics are reported to account for class imbalance, which is particularly pronounced in the target-identification setting where the Individual class is underrepresented relative to the Group class.

4. Results

The performance of the proposed mBERT-based framework was evaluated across multilingual social support detection tasks using accuracy, macro-averaged F1-score, and weighted F1-score. Due to class imbalance within the datasets, macro F1-score was considered particularly important for assessing balanced classification performance across minority and majority classes.

Table 3 and 4 summarize the classification performance of the fine-tuned `bert-base-multilingual-cased` (mBERT) model across all evaluated subtasks and languages. All experiments report precision, recall, F1-score, and overall accuracy on the held-out test sets.

Table 3
Classification results of mBERT for the support detection Task 1 across English and Spanish datasets.

Language	Class	Precision	Recall	F1-score	Support
English	Non-Support	0.8956	0.9741	0.9332	502
English	Support	0.8762	0.6174	0.7244	149
Spanish	Non-Support	0.9069	0.9185	0.9127	1559
Spanish	Support	0.6983	0.6667	0.6821	441

Table 4
Classification results of mBERT for the support target classification Subtask 2 across English and Spanish datasets.

Language	Class	Precision	Recall	F1-score	Support
English	Group	0.7932	0.6446	0.7112	363
English	Individual	0.5556	0.2564	0.3509	78
Spanish	Group	0.6837	0.5537	0.6119	121
Spanish	Individual	0.7917	0.5278	0.6333	36

Table 5 presents the rankings achieved by our system for the officially reported Subtasks 1 and 2 across the English and Spanish tracks in SSD-2026. The proposed approach demonstrated stronger performance on the English datasets, obtaining 15th place in Task 1 and 14th place in Task 2. In comparison, the rankings for the Spanish datasets were 40th in Task 1 and 38th in Task 2. Since Subtask 3 was not included in the reported official submission, no ranking is reported for that subtask. The difference in rankings may be attributed to variations in dataset characteristics, linguistic complexity, class distributions, or language-specific challenges encountered during model training and evaluation.

Table 5
Ranking achieved in the shared task for English and Spanish tracks.

Language	Task	Rank
Spanish	Task 1	40
Spanish	Task 2	38
English	Task 1	15
English	Task 2	14

4.1. Subtask 1: Support Detection

As shown in Table 1, the proposed mBERT model achieved strong performance for Task 1 across both English and Spanish datasets. The Spanish dataset produced the highest macro-averaged F1-score of 82.88%, while English obtained 79.74%. These results indicate reasonably balanced classification performance between supportive and non-supportive categories, although class imbalance still affected recall for supportive comments.

For the binary support detection task, mBERT achieved an accuracy of 86.30% with a macro-F1 of 0.7974 on English, and an accuracy of 89.25% with a macro-F1 of 0.8288 on Spanish. In both languages, the Non-Supportive class achieved substantially higher F1-scores than the Supportive class, showing that the model tends to under-predict supportive comments.

4.2. Subtask 2: Support Target Classification

For the binary support target classification task, which distinguishes between Group-directed and Individual-directed support, mBERT achieved an accuracy of 87.30% on English and 86.02% on Spanish. In both languages, the model performed better on the majority Group class than on the minority Individual class. Macro-averaged F1-scores of 0.6622 for English and 0.7240 for Spanish show that class imbalance affected balanced performance, especially for individually directed support. The results indicate that target identification remains more challenging than binary support detection.

4.3. Cross-lingual Observations

Across both subtasks, mBERT demonstrated broadly comparable performance between English and Spanish, with accuracy differences of no more than three percentage points. This is consistent with the model's multilingual pre-training, which enables effective transfer across languages without task-specific language adaptation. However, the consistently lower recall observed for minority classes, particularly *Individual* in Spanish Subtask 2, suggests that class imbalance remains a significant challenge that future work could address through oversampling, class-weighted loss functions, or data augmentation strategies.

5. Discussion

The experimental results demonstrate that fine-tuning `bert-base-multilingual-cased` on the IberLEF 2026 Social Support Detection dataset yields competitive performance across both languages and subtasks. Several observations merit further discussion.

5.1. Performance on Support Detection (Subtask 1)

mBERT achieved its highest overall accuracy on Spanish Subtask 1 (89.25%), while English Subtask 1 reached 86.30%. Binary support detection is more tractable than target identification, because the model reliably identifies non-supportive comments in both languages, with F1-scores above 0.91. However, the recall for the Supportive class remains lower (0.6174 in English and 0.6667 in Spanish), indicating that the model is conservative in predicting support, likely due to class imbalance and the implicit nature of supportive language in social media.

5.2. Performance on Support Target Classification (Subtask 2)

The binary target classification task proved more challenging, particularly for the minority Individual class. In both English and Spanish, Group-directed support was easier to identify than individually directed support. The Individual class achieved lower recall and F1, especially in English, where its F1-score was 0.3509. This difficulty is likely related to the smaller number of Individual-labelled samples and the varied ways users address specific people in informal social media comments.

5.3. Cross-lingual Transfer

A key strength of the proposed approach is its ability to handle both English and Spanish within a single model architecture, without language-specific fine-tuning. The comparable performance across languages, with accuracy differences of under three percentage points on both subtasks, confirms that mBERT's multilingual representations are effective for cross-lingual social support detection. This is particularly valuable for low-resource settings where monolingual annotated data may be scarce.

5.4. Limitations

Despite its strengths, this study has several limitations. First, class imbalance across subtasks was not explicitly addressed during training; techniques such as class-weighted loss, oversampling of minority classes, or focal loss could potentially improve recall for underrepresented categories. Second, the reported official results cover Subtasks 1 and 2 only; Subtask 3 should be investigated in future work to provide a complete evaluation of community-level support classification. Third, the current approach does not incorporate community-specific or domain-adapted pre-training, which may limit sensitivity to nuanced online social support language. Future work could explore models such as BERTweet for English or Spanish-specific transformers such as BETO.

6. Conclusion

This paper presented a transformer-based approach to multilingual social support detection using the IberLEF 2026 shared task dataset. We fine-tuned mBERT for binary support detection (Subtask 1) and support target classification (Subtask 2) across English and Spanish comments. The model achieved strong overall accuracy on the reported subtasks, ranging from 86.02% to 89.25%, with robust performance on the dominant Non-Supportive and Group classes.

The results show that multilingual transformers provide an effective baseline, but class imbalance remains a key bottleneck, especially for Individual-directed support in Spanish. Future work should explore data augmentation, class-weighted training, ensemble strategies, and a complete evaluation of Subtask 3 for community-level support classification.

Overall, this study contributes to NLP research on prosocial communication and shows that transformer-based models can support scalable analysis of supportive interactions across languages and platforms.

Declaration on Generative AI

During the preparation of this work, the authors reviewed and edited the content by using generative AI tools to assist with language editing (grammar and clarity) and to assist in debugging and refactoring portions of the experimental code. After using these tools, the authors carefully reviewed and edited all generated content as necessary and take full responsibility for the content of this publication.

7. Acknowledgments

The authors would like to thank the organizers of the SSD-2026 Shared Task at IberLEF 2026 for providing the dataset and evaluation framework. The first and second authors acknowledge the support of the Instituto Politécnico Nacional (IPN) through the Centro de Investigación en Computación (CIC). Computational resources were provided via Google Colab.

References

[1] B. H. Kaplan, J. C. Cassel, S. Gore, Social support and health, Medical Care 15 (1977) 47–58.

[2] R. L. Leavy, Social support and psychological disorder: A review, J. Community Psychol. 11 (1983) 3–21.

[3] A. G. Billings, R. H. Moos, Coping, stress, and social resources among adults with unipolar depression, J. Pers. Soc. Psychol. 46 (1984) 877–891.

[4] W. Stroebe, M. Stroebe, G. Abakoumkin, H. Schut, The role of loneliness and social support in adjustment to loss: A test of attachment versus stress theory, J. Pers. Soc. Psychol. 70 (1996) 1241–1249.

[5] S. Cohen, L. G. Underwood, B. H. Gottlieb, Social Support Measurement and Intervention: A Guide for Health and Social Scientists, Oxford University Press, New York, NY, 2000.

[6] D. N. Akinloye, Communication and Online Platforms: Pitfalls, Possibilities, and Pressures, Ph.D. thesis, 2024.

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

[8] M. C. Stewart, B. G. Wilson, The dynamic role of social media during hurricane Sandy: An introduction of the STREMII model to weather the storm of the crisis lifecycle, Computers in Human Behavior 54 (2016) 639–646.

[9] J. Huh, M. S. Ackerman, Collaborative help in chronic disease management: Supporting individualized problems, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, 2012, pp. 853–862.

[10] S. Thomas, P. Kanske, J. Schäfer, K. V. Hummel, S. Trautmann, Examining bidirectional associations between perceived social support and psychological symptoms in the context of stressful event exposure: A prospective, longitudinal study, BMC Psychiatry 22 (2022) 736.

[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] S. B. Låftman, A. Lundin, V. Östberg, Perceived social support and symptoms of depression and anxiety in emerging adulthood: A Swedish prospective cohort study, Scandinavian Journal of Public Health 53 (2025) 854–862.

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

[14] Y. Ma, Z. Zhou, C. Ye, M. Liu, Online social support and problematic internet use: A meta-analysis, Addictive Behaviors 160 (2025) 108160. doi:10.1016/j.addbeh.2024.108160.

[15] V. I. Erčulj, A. Žiberna, The role of online social support in patients undergoing infertility treatment–a comparison of pregnant and non-pregnant members, Health communication 37 (2022) 1724–1730.

[16] 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).

[17] M. Kerasiotis, L. Ilias, D. Askounis, Depression detection in social media posts using transformer-based models and auxiliary features, Social Network Analysis and Mining 14 (2024) 196.

[18] M. Zamir, M. Tash, Z. Ahani, A. Gelbukh, G. Sidorov, Tayyab@ dravidianlangtech 2024: detecting fake news in malayalam lstm approach and challenges, in: Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, 2024, pp. 113–118.

[19] B. G. Bokolo, Q. Liu, Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts, Electronics 13 (2024) 3980.

[20] M. T. Zamir, F. Ullah, R. Tariq, W. H. Bangyal, M. Arif, A. Gelbukh, Machine and deep learning algorithms for sentiment analysis during covid-19: A vision to create fake news resistant society, PloS one 19 (2024) e0315407.

[21] 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 (2026). URL: https://doi.org/10.21203/rs.3.rs-9557287/v1. doi:10.21203/rs.3.rs-9557287/v1.

[22] Z. Ahani, M. S. Tash, M. Zamir, I. Gelbukh, A. Gelbukh, Zavira@ dravidianlangtech 2024: Telugu hate speech detection using lstm, in: Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, 2024, pp. 107–112.

[23] P. Broadwell, N. Davis, S. Yoon, Using artificial intelligence to develop a lexicon-based African American tweet detection algorithm to inform culturally sensitive Twitter-based social support interventions for African American dementia caregivers, in: Informatics and Technology in Clinical Care and Public Health, SAGE Publications, London, UK, 2022, pp. 1–4.

[24] P. Mohaghegh, N. Roozbahani, K. Vakilian, M. Radpour, et al., Relationship of perceived social support with health-promoting lifestyle in women participating in national breast cancer early detection program, Iranian Journal of Psychiatry and Clinical Psychology 26 (2021) 464–477.

[25] T. Achamaleh, F. Uroosa, N. Hafeez, T. O. Abiola, M. Mebraihtu, S. Getachew, G. Sidorov, R. Quintero, EM-26@LT-EDI 2025: Detecting racial hoaxes in code-mixed social media data, in: K. Gkirtzou, S. Žitnik, J. Gracia, D. Gromann, M. P. di Buono, J. Monti, M. Ionov (Eds.), Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, Unior Press, Naples, Italy, 2025, pp. 146–152.

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

[27] F. Uroosa, A. Abbas, M. T. Zamir, G. Sidorov, Speech-based emotion recognition and classification integrating a CNN and BiLSTM network, in: Proceedings of the Language Resources and Evaluation Conference (LREC-2026): Workshop on Computational Affective Science, 2026.