Multilingual Social Support Detection Using Transformer-Based Sentence Embeddings and Hierarchical Classification at IberLEF 2026

Sara Barahouei Nouri[1], Abdollah Abadian[2,*], Ismael Jacinto Dias Meque[3], Grigori Sidorov[2,*] and Olga Kolesnikova[2]

[1]Islamic Azad University Zahedan Branch, Zahedan, Iran
[2]Center for Computing Research, Instituto Politécnico Nacional (IPN), Mexico City 07320, Mexico
[3]Catholic University of Mozambique, Mozambique

Abstract

This paper presents our submission to the Social Support Detection (SSD) Shared Task at IberLEF 2026, which focuses on identifying supportive language, its intended targets, and specific supported communities across English and Spanish social-media comments. We conducted a systematic evaluation of six progressively advanced machine-learning architectures, transitioning from traditional lexical baselines (TF-IDF with LinearSVC) to dense multilingual semantic representations (MiniLM, LaBSE, and MPNet) combined with nonlinear Support Vector Machines (SVMs). To address the hierarchical nature of the subtasks, we implemented a deterministic post-processing strategy that enforces logical consistency across predictions. Our best-performing system uses multilingual MPNet embeddings with an optimized RBF-kernel SVM. On the English dataset, our best run achieved Macro-F1 scores of 0.8070, 0.6969, and 0.4590 for Tasks 1–3, ranking 13th, 28th, and 20th out of 40 teams respectively. On the Spanish dataset, the same system achieved Macro-F1 scores of 0.8445, 0.8146, and 0.5273, ranking 17th, 20th, and 30th out of 41 teams respectively. While our models effectively captured general supportive intent, performance decreases on fine-grained target and community classification highlight the need for deeper contextual and sociocultural reasoning in future research.

Keywords

Social Support Detection, Multilingual NLP, Sentence Embeddings, Transformer Models, MPNet, Support Vector Machines, Social Media Analysis, Hierarchical Classification, IberLEF 2026

1. Introduction

Social support—encompassing emotional comfort, informational guidance, appraisal validation, and instrumental assistance—plays a fundamental role in promoting psychological well-being, reducing stress, and fostering resilience among individuals facing adversity [1, 2]. With the proliferation of social-media platforms, online communities have emerged as accessible, anonymous, and scalable venues where individuals seek and provide support, particularly for sensitive mental-health concerns such as depression, anxiety, substance-use disorders, and trauma [3, 4]. Unlike traditional offline support networks, online platforms enable real-time, cross-cultural, and multilingual interactions, creating both opportunities and challenges for automated social support detection.

1.1. Social Support Detection as an NLP Task

Ahani et al. [5] introduced the Social Support Detection (SSD) task as a hierarchical three-subtask problem: (1) binary support detection (Support / Not Support), where supportive comments express encouragement, care, admiration, help, or solidarity and comments promoting violence are labeled Not

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
barahoueisarah@gmail.com (S. B. Nouri); aabdollaha26@cic.ipn.mx (A. Abadian); 701210417@ucm.ac.mz (I. J. D. Meque); sidorov@cic.ipn.mx (G. Sidorov); kolesnikova@cic.ipn.mx (O. Kolesnikova)
0009-0002-7747-1612 (S. B. Nouri); 0009-0004-6581-4053 (A. Abadian); 0009-0003-4202-8159 (I. J. D. Meque); 0000-0003-3901-3522 (G. Sidorov); 0000-0002-1307-1647 (O. Kolesnikova)

Support; (2) binary target identification (Individual / Group); and (3) multiclass group identification (Black Community, LGBTQ, Nation, Other, Religion, Women), evaluated separately for English and Spanish.

Subsequent work extended this framework to Spanish-language content. Shahiki Tash et al. [6] created the first annotated Spanish dataset of 3,189 YouTube comments, revealing significant class imbalance (only 679 supportive comments). Their experiments showed that balanced datasets (via GPT-4o paraphrasing) improved Tasks 2 and 3, while GPT-4o achieved the best Task-1 macro F1 of 0.8531.

Kolesnikova et al. [7] advanced the state of the art by applying transformer-based models and zero-shot learning with GPT-3, GPT-4, and GPT-4o. RoBERTa-base consistently outperformed other models, achieving macro F1 scores of 0.78, 0.84, and 0.80 on an imbalanced dataset. He et al. [8] demonstrated that incorporating social support information into dialogue systems significantly improved recognition performance, achieving F1 of 0.879 in Mobile Instant Messaging Health Communities.

Cross-linguistic differences remain underexplored. Shahiki-Tash et al. [9] conducted a comparative LIWC-based analysis of English and Spanish supportive comments, finding that Spanish contexts exhibit significantly higher appraisal support ($p < 0.001$), underscoring the importance of culture-specific models.

1.2. Research Gap and Contributions

Despite these advances, several critical gaps remain. First, most existing studies focus on a single language. Second, systematic comparisons across English and Spanish datasets are limited. Third, the impact of class imbalance on fine-grained community classification remains underexplored. This paper makes the following contributions:

- **Systematic evaluation of multilingual sentence embeddings:** six experimental systems progressing from TF-IDF baselines to transformer-based semantic representations (MiniLM, LaBSE, MPNet, multilingual MPNet) with RBF-SVM classifiers.
- **Hierarchical task framework:** a deterministic post-processing strategy enforcing logical consistency across all three subtasks.
- **Cross-lingual performance analysis:** detailed comparative results for English and Spanish, identifying task difficulty hierarchies and language-specific challenges.
- **Reproducible lightweight pipeline:** our best system achieves macro F1 of 0.8445 (Spanish Task 1) without computationally expensive fine-tuning.

2. Related Work

2.1. Multilingual and Code-Mixed Text Classification

Balla Venkata et al. [10] addressed emotion recognition in English–Roman Urdu code-mixed tweets using an enhanced MBART-CM framework, achieving 80% accuracy. Tesfagergish et al. [11] proposed a semi-supervised method for sentiment analysis achieving 87.3% accuracy on SemEval-2017. Their follow-up [12] studied Amharic sentiment classification, finding that a Feed-Forward Neural Network combined with sentence transformers achieved 82.2% accuracy.

2.2. Classical and Generative Approaches

Mahmood et al. [13] demonstrated that SBERT-based features with a LinearSVC significantly outperformed generative models, achieving Micro-F1 of 0.9142 for code-mixed emotion classification. Usman et al. [14] developed a multilingual hate-speech detection framework where GPT-3.5-turbo outperformed XLM-R by an average 4% across English, Spanish, and Urdu. Tharunika et al. [15] proposed a zero-shot, translation-free emotion detection framework using LaBSE and adaptive emotion prototypes.

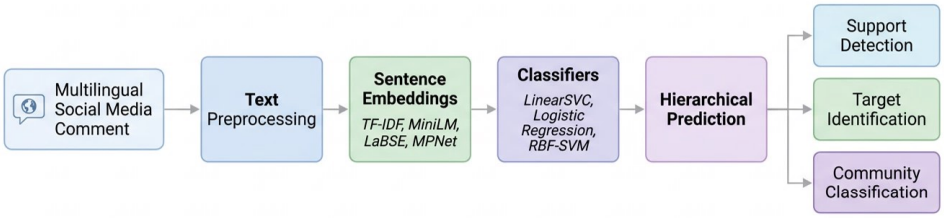

Figure 1: Overall workflow of the proposed multilingual social support detection framework.

Petkevičius and Vitkutė-Adžgauskienė [16] demonstrated that data augmentation via neural machine translation could effectively fine-tune transformer models for emotion recognition, achieving 81.3% accuracy.

2.3. Benchmarks and Surveys

Al Maruf et al. [17] surveyed text-based emotion detection, identifying dataset imbalance and domain adaptation as key challenges. Guo et al. [18] benchmarked transformer-based models on 25 social-media datasets, showing that RoBERTa-base and BERTweet outperform clinical models for social-media tasks. Dehghan and Yanikoglu [19] proposed BERTurk-DualCL, achieving 81.04% macro F1 for Turkish hate speech. Taranis et al. [20] found that GPT-4o achieved the highest overall F1 (0.84) for multilingual detection of irregular migration discourse.

3. Research Methodology

3.1. Overview

This work investigates multilingual approaches for social support detection in the SSD Shared Task at IberLEF 2026, introduced in the SSD-2026 overview paper and organized as part of IberLEF 2026 [21, 22]. The proposed systems address three hierarchical subtasks: (1) support detection, (2) target identification, and (3) community-level multiclass classification.

Our methodology consists of four main stages: text preprocessing, sentence-embedding extraction, supervised classification, and hierarchical post-processing (Figure 1). Multiple experimental runs evaluated the impact of different embedding models and classifiers on both English and Spanish datasets.

3.2. Data Preprocessing

The official training and test datasets were used for all experiments [5, 7, 6]. The training set consists of 7,998 comments and the test set of 2,000 comments, with approximately equal representation of English

and Spanish samples.

Social-media comments were cleaned and normalised before training. All text was lower-cased; URLs, hyperlinks, user mentions, and unnecessary special symbols were removed using regular expressions; extra spaces and noisy characters were eliminated; and missing entries were replaced with empty strings.

3.3. Text Representation

3.3.1. TF-IDF Baseline

Comments were represented as Term Frequency–Inverse Document Frequency (TF-IDF) vectors and paired with a Linear Support Vector Classifier (LinearSVC) to establish a lexical baseline.

3.3.2. Transformer-Based Sentence Embeddings

In later runs, several pre-trained multilingual sentence-embedding models were evaluated: `paraphrase-multilingual-MiniLM-L12-v2`, `LaBSE`, `all-mpnet-base-v2`, and `paraphrase-multilingual-mpnet-base-v2`. The multilingual MPNet and LaBSE models showed the strongest capability in projecting both English and Spanish comments into a shared semantic space.

3.4. Classification Models

Separate classifiers were trained independently for each subtask, allowing task-specific decision boundaries without parameter interference.

3.4.1. LinearSVC

Used with TF-IDF features as a strong lexical baseline.

3.4.2. Logistic Regression

Used with MiniLM embeddings for Run 2; selected for computational efficiency with dense representations.

3.4.3. Nonlinear RBF-SVM

Runs 3–6 used an SVM with a Radial Basis Function (RBF) kernel. Input embeddings were normalized with `StandardScaler`. Class imbalance was addressed with `class_weight='balanced'`. Regularisation parameter values of $C = 10$, $C = 15$, and $C = 20$ were tested; gamma was set to `'scale'` (i.e., $1/(\text{n\_features} \times \text{Var}(X))$) throughout.

3.5. Hierarchical Prediction Strategy

The subtasks follow a dependency structure: a comment classified as Not Support should not have a valid support target or community category. A rule-based post-processing step enforces this: when Task 1 predicts Not Support, Task 2 is set to Group (label 0) and Task 3 to Other (label 3). This reduces contradictory outputs and improves evaluation stability.

3.6. Experimental Runs

Table 1 summarises the six submitted systems.

Difference between Run 5 and Run 6. Both use `paraphrase-multilingual-mpnet-base-v2` embeddings with class-balanced RBF-SVM. The only difference is the regularisation parameter: Run 5 uses $C = 15$, while Run 6 uses $C = 20$. The higher C in Run 6 was intended to reduce misclassification

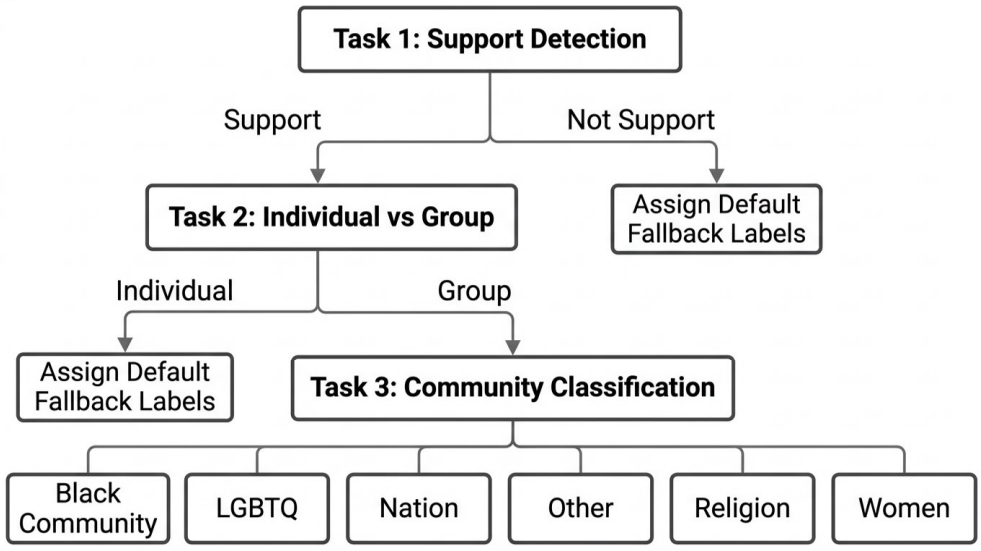

Figure 2: Hierarchical dependency structure used for enforcing prediction consistency across subtasks.

Table 1
Summary of experimental runs.

Run	Representation	Classifier
Run 1	TF-IDF	LinearSVC
Run 2	MiniLM Embeddings	Logistic Regression
Run 3	LaBSE Embeddings	RBF-SVM ($C = 10$)
Run 4	MPNet English Embeddings	RBF-SVM ($C = 10$)
Run 5	Multilingual MPNet Embeddings	RBF-SVM ($C = 15$)
Run 6	Multilingual MPNet (enhanced)	RBF-SVM ($C = 20$)

on minority classes by allowing smaller margins. Both runs achieved an identical Macro-F1 of 0.8445 on the official leaderboard (Task 1 binary detection), indicating that the decision boundary was already near-optimal at $C = 15$ and that increasing C further did not yield measurable gains on this metric.

3.7. Reproducibility Details

To facilitate replication of our experiments, the complete implementation details are as follows. All code was written in Python 3.10 and executed on Google Colab with NVIDIA T4 GPU acceleration. Sentence embeddings were generated using the `sentence-transformers` (v2.2.2) library; classifiers were trained with `scikit-learn` (v1.2.2). A stratified 80/20 split of the training set was used for hyperparameter selection: 6,398 comments for training and 1,600 for validation. The random seed was fixed to 42 for all splits, SVM solvers, and embedding generation where applicable. Table 2 lists the final hyperparameters for each run.

3.8. Evaluation Metric

The official metric was Macro-F1 score, which equally weights all classes and is particularly appropriate for the imbalanced label distributions in Tasks 2 and 3. Performance was evaluated independently for each subtask and each language.

Table 2
Hyperparameter details for all experimental runs.

Run	Classifier	C	Gamma	class_weight
Run 1	LinearSVC	1.0	—	balanced
Run 2	Logistic Regression	1.0	—	balanced
Run 3	RBF-SVM	10	scale	balanced
Run 4	RBF-SVM	10	scale	balanced
Run 5	RBF-SVM	15	scale	balanced
Run 6	RBF-SVM	20	scale	balanced

Table 3
Performance of submitted systems on binary support detection (Task 1 official leaderboard score).

Run	Representation	Classifier	Macro-F1
Run 1	TF-IDF	LinearSVC	0.7826
Run 2	MiniLM Embeddings	Logistic Regression	0.7575
Run 3	LaBSE Embeddings	RBF-SVM	0.8070
Run 4	MPNet English Embeddings	RBF-SVM	0.8069
Run 5	Multilingual MPNet	RBF-SVM ($C = 15$)	0.8445
Run 6	Multilingual MPNet (enhanced)	RBF-SVM ($C = 20$)	0.8445

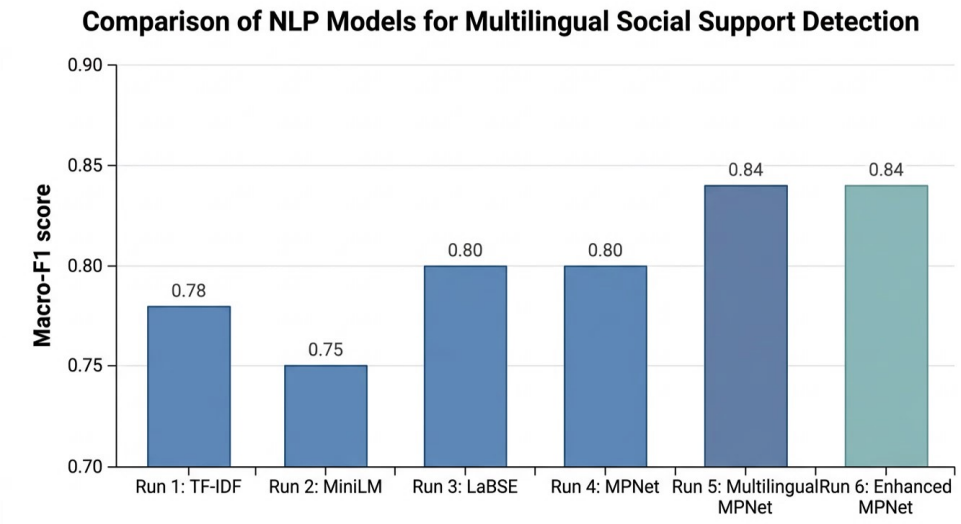

Figure 3: Macro-F1 comparison across the six evaluated systems.

4. Results and Discussion

4.1. Overall Results

Table 3 presents the Macro-F1 scores of all six submitted systems on the official leaderboard (Task 1, binary support detection, combined). Performance improves monotonically from Run 1 to Run 5, confirming the value of richer semantic representations and nonlinear classifiers.

The TF-IDF baseline (Run 1: 0.7826) already provides a competitive result, demonstrating the power of well-tuned lexical baselines. The MiniLM system (Run 2: 0.7575) under-performed the TF-IDF baseline,

Table 4
Spanish leaderboard performance (Run 5/6, best system).

Subtask	Macro-F1	Rank (of 41)
Task 1: Binary Support Detection	0.8445	17^{th}
Task 2: Target Identification	0.8146	20^{th}
Task 3: Community Classification	0.5273	30^{th}

Table 5
English leaderboard performance (Run 5/6, best system).

Subtask	Macro-F1	Rank (of 40)
Task 1: Binary Support Detection	0.8070	13^{th}
Task 2: Target Identification	0.6969	28^{th}
Task 3: Community Classification	0.4590	20^{th}

Table 6
Combined English and Spanish performance of the best system (Runs 5 and 6) across all three subtasks.

	English			Spanish		
Subtask	Macro-F1	Rank	N_{teams}	Macro-F1	Rank	N_{teams}
Task 1: Support Detection	0.8070	13	40	0.8445	17	41
Task 2: Target Identification	0.6969	28	40	0.8146	20	41
Task 3: Community Class.	0.4590	20	38	0.5273	30	39

suggesting that smaller embedding models with linear decision boundaries are insufficient for the complex semantics of social support. Systems using RBF-SVM consistently outperformed Logistic Regression, indicating that the decision boundary between supportive and non-supportive language is highly nonlinear. The best-performing multilingual MPNet systems (Runs 5 and 6: 0.8445) represent a 7.9 percentage-point improvement over the TF-IDF baseline.

4.2. Spanish Leaderboard Results

Table 4 summarises the official leaderboard results for Spanish subtasks.

Our system achieved its strongest Spanish performance on Task 1 (binary support detection). Performance decreased on Task 2, where support targets are often implicit or context-dependent, requiring deeper pragmatic understanding. The sharpest drop occurred on Task 3, which requires fine-grained sociocultural distinctions across six community categories.

4.3. English Leaderboard Results

Table 5 presents the official English leaderboard performance.

Our system ranked 13th on English Task 1, demonstrating competitive binary classification. However, Task 2 dropped to rank 28th and Task 3 to 20th. English social-media discourse frequently contains implicit references, abbreviated expressions, and ambiguous target structures, which challenge fine-grained classification.

4.4. Cross-Language and Cross-Task Analysis

Table 6 provides a combined view of performance across both languages and all tasks for our best system (Run 5/6).

Two cross-language patterns stand out. First, Spanish performance consistently exceeds English performance: the gap is 3.75 pp on Task 1, 11.77 pp on Task 2, and 6.83 pp on Task 3. This may reflect

the richer training signal available in the Spanish portion of the dataset, or greater homogeneity in Spanish social-media support expressions. Second, the difficulty hierarchy (Task 1 $\gg$ Task 2 $>$ Task 3) holds for both languages: binary detection is the most tractable subtask, while fine-grained community classification is the most challenging.

4.5. Discussion

Effectiveness of semantic embeddings. The progression from Run 1 to Run 5 confirms that multilingual sentence embeddings substantially outperform lexical representations. The multilingual MPNet model—trained on parallel corpora across 50+ languages—provides a shared semantic space that is particularly effective for cross-lingual social support detection.

Run 5 vs. Run 6. As noted in Section 3.5, both runs achieved identical Macro-F1 (0.8445 on Task 1). The null effect of increasing C from 15 to 20 suggests that the SVM was already well-calibrated at $C = 15$. Future tuning could explore the gamma parameter more aggressively, or combine multiple classifiers in an ensemble.

Community classification challenges. The lowest performance on Task 3 (0.4590 English, 0.5273 Spanish) reflects the inherent difficulty of distinguishing among six community categories (Black Community, LGBTQ, Nation, Other, Religion, Women). Many comments refer to groups indirectly. Static sentence embeddings lack the sociocultural grounding needed to resolve such ambiguities. Future work should explore entity-aware representations or community-specific fine-tuning.

Limitations. Our approach does not include any fine-tuning of the transformer encoder, ensemble methods, or data augmentation. The 80/20 validation split was used only for C selection; broader cross-validation could yield more robust hyperparameters.

5. Conclusion

This paper presented our submission to the Social Support Detection Shared Task at IberLEF 2026, addressing the hierarchical classification of supportive language, its intended targets, and specific supported communities across English and Spanish social-media comments. Through a systematic evaluation of six progressively advanced architectures, we demonstrated that dense multilingual semantic representations substantially outperform lexical baselines for capturing implicit markers of social support.

Our best-performing system, leveraging multilingual MPNet embeddings with an optimized RBF-SVM, achieved Macro-F1 scores of 0.8070 (English Task 1, rank 13/40) and 0.8445 (Spanish Task 1, rank 17/41). The deterministic hierarchical post-processing strategy proved effective at reducing contradictory predictions. A notable finding is that Run 5 ($C = 15$) and Run 6 ($C = 20$) achieved identical Macro-F1 scores, indicating that the decision boundary was already near-optimal at $C = 15$.

Despite competitive Task-1 results, performance on Tasks 2 and 3 remains substantially lower, with community classification being the most challenging. Future work should focus on: (1) end-to-end fine-tuning of multilingual transformers; (2) ensemble methods; (3) data augmentation for minority community classes; and (4) explicit modelling of sociocultural context. Despite current limitations, our results confirm that lightweight multilingual sentence-embedding pipelines provide a strong and reproducible baseline for social support detection.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) for grammar checking, phrasing suggestions, and minor language polishing of the manuscript. GitHub Copilot was used for

code completion suggestions during implementation of the SVM models and data preprocessing scripts. After using these tools, the authors reviewed and edited all content as needed and take full responsibility for the publication's content. No generative AI was used to produce experimental results, to generate predictions, or to perform data annotation.

Acknowledgments

The authors thank the SSD-2026 task organizers for providing the dataset and evaluation infrastructure.

References

[1] F. Xia, et al., Social Support and Online Communities, Journal of Computer-Mediated Communication (2012).

[2] C. P. H. Langford, J. Bowsher, J. P. Maloney, P. P. Lillis, Social Support: A Conceptual Analysis, Journal of Advanced Nursing 25 (1997) 95–100.

[3] M. De Choudhury, S. De, Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity, in: Proceedings of the ICWSM, 2014.

[4] N. Andalibi, P. Ozturk, A. Forte, Social Support and Discord in Mental Health Online Communities, in: Proceedings of the ACM CHI Conference, 2018.

[5] Z. Ahani, M. S. Tash, F. Balouchzahi, L. Ramos, G. Sidorov, A. Gelbukh, R. Monroy, Social Support Detection from Social Media Texts, PLOS ONE 21 (2026) e0337476. doi:10.1371/journal.pone.0337476.

[6] 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).

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

[8] C. He, F. Liu, Q. Huang, J. Wu, Social-Support-Aware Dialogue Recognition in Mobile Instant Messaging Online Health Communities, Information Technology and Management (2025). doi:10.1007/s10799-026-00485-6.

[9] M. Shahiki-Tash, Z. Ahani, L. Ramos, O. Kolesnikova, A. Y. Barrera-Animas, Linguistic Analysis in Different Types of Support Using LIWC for Spanish and English Online Communities, 2026. https://doi.org/10.21203/rs.3.rs-9557287/v1.

[10] S. Balla Venkata, A. Alzubaidy, S. Kumar M., Emotion Recognition in Code-Mixed Social Media Data Using an Optimized Transformer Architecture, in: 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), IEEE, 2025.

[11] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, R. Damaševičius, Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning, Applied Sciences 12 (2022) 8662.

[12] S. G. Tesfagergish, R. Damaševičius, J. Kapočiūtė-Dzikienė, Deep Learning-Based Sentiment Classification in Amharic Using Multi-Lingual Datasets, Computer Science and Information Systems 20 (2023) 1459–1481.

[13] A. Mahmood, M. Torres-Ruiz, Z. Ahmad, H. Farid, I. Ameer, R. Quintero, An Efficient Approach for Code-Mixed Emotion Classification Applying Machine Learning, IEEE Access 13 (2025).

[14] M. Usman, M. Ahmad, G. Sidorov, I. Gelbukh, R. Q. Tellez, A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media, Computers 14 (2025) 279.

[15] L. Tharunika, S. Rajasulochana, T. Shanmugapriya, K. Sangeetha, P. Poonkodi, Zero-Shot Cross-Cultural Emotion Detection Using Language-Agnostic Embeddings and Adaptive Emotion Prototypes, in: 2025 International Conference on Next Generation Computing Systems (ICNGCS), IEEE, 2025.

[16] M. Petkevičius, D. Vitkutė-Adžgauskienė, Fine-Tuning Language Models for Emotion Recognition in Lithuanian Texts Using Neural Machine Translation of Training Datasets, CEUR Workshop Proceedings (2023).
[17] A. Al Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, Z. Aung, Challenges and Opportunities of Text-Based Emotion Detection: A Survey, IEEE Access 12 (2024) 18430–18450.
[18] Y. Guo, X. Dong, M. A. Al-Garadi, A. Sarker, C. Paris, D. Mollá-Aliod, Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets, in: Proceedings of the Australasian Language Technology Association Workshop, 2020.
[19] S. Dehghan, B. Yanikoglu, Multi-Domain Hate Speech Detection Using Dual Contrastive Learning and Paralinguistic Features, Technical Report, Sabanci University (2022).
[20] D. Taranis, G. Razis, I. Anagnostopoulos, A Pilot Study on Multilingual Detection of Irregular Migration Discourse on X and Telegram Using Transformer-Based Models, Electronics 15 (2026) 281.
[21] 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.
[22] 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.