OptimAlzers at Social Support Detection 2026: Transformer Models for Social Support Detection in Spanish and English Social Media Comments

Emmanuel Castro[1,†], Irari Jiménez-López[1,†], Yoqsan Angeles[1,†], José E. Valdez-Rodríguez[1], Olga Kolesnikova[1,*], Hiram Calvo[1] and Carlos Aguilar-Ibañez[1]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Abstract

Social media platforms have become important spaces for expressing opinions, emotions, and diverse forms of social interaction. The automatic detection of supportive and non-supportive comments can contribute to the analysis of online behavior, the identification of vulnerable communities, and the development of tools for monitoring healthier digital environments. Motivated by this context, this work presents an experimental pipeline for classifying English and Spanish YouTube comments using state-of-the-art transformer-based models. The study was conducted within the Social Support Detection 2026 shared task at IberLEF 2026, which includes three subtasks: binary classification of supportive and non-supportive comments, binary classification of comments directed at individuals or groups, and multi-class classification of comments to a specific community, including categories such as nationality, LGBTQ community, Black community, women, and religion. Five transformer models were fine-tuned and evaluated for each task. For English comments, the best models achieved an F1-score of 0.80, 0.84, and 0.87 for Tasks 1, 2, and 3, respectively. For Spanish comments, the best F1-score were 0.82, 0.84, and 0.57, respectively. These results suggest that transformer-based models can effectively identify social support patterns in multilingual YouTube comments. In addition, the findings indicate that language-specific models may offer advantages over multilingual models in this task.

Keywords

Social Support, Spanish, Transformers, Fine-Tuning, NLP, Classification

1. Introduction

With the increasing reach of the Internet and social media platforms, the dissemination of harmful content has become a major societal concern [1]. As a result, a substantial body of research in Natural Language Processing (NLP) has focused on detecting, moderating, and filtering negative forms of online interaction, such as hate speech, offensive language, or misinformation.

However, online communication is not limited to harmful or antisocial content. Social media platforms also enable users to express empathy, encouragement, solidarity, advocacy, and emotional assistance toward individuals and communities facing different social, political, or personal challenges. Despite its relevance, supportive interaction remains comparatively less explored than other content categories, like harmful content detection. This lack of research suggests the need to design computational strategies capable not only of identifying negative content, but also of recognizing socially beneficial forms of communication.

Online social support can be understood as the emotional, informational, or symbolic assistance provided through digital platforms. This type of support is particularly relevant for individuals and

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

ccastrom2023@cic.ipn.mx (E. Castro); ijimenezl2024@cic.ipn.mx (I. Jiménez-López); yangelesg2020@cic.ipn.mx (Y. Angeles); jvaldezr2018@cic.ipn.mx (J. E. Valdez-Rodríguez); kolesnikova@cic.ipn.mx (O. Kolesnikova); hcalvo@cic.ipn.mx (H. Calvo); carlosaguilari@cic.ipn.mx (C. Aguilar-Ibañez)

0009-0006-0297-0563 (E. Castro); 0009-0007-1171-3435 (I. Jiménez-López); 0009-0004-0886-5540 (Y. Angeles); 0000-0002-4572-5713 (J. E. Valdez-Rodríguez); 0000-0002-8372-4228 (O. Kolesnikova); 0000-0003-2836-2102 (H. Calvo); 0000-0003-3925-2435 (C. Aguilar-Ibañez)

groups exposed to vulnerability, discrimination, conflict, or crisis situations, such as marginalized communities, or people seeking recognition. Detecting supportive content is therefore important for understanding how solidarity emerges in online spaces so it can promote healthier and more constructive digital environments [2, 3, 4].

This study focuses on identification of support in online comments instead of focusing on the detection of harmful content. By analyzing supportive expressions in Spanish and English language YouTube comments, this work contributes to the development of NLP resources and methods aimed at characterizing positive social interaction.

Recent work has begun to formalize Social Support Detection (SSD) as a specific NLP task. Ahani et al. [5] introduced SSD as the automatic identification of supportive interactions within online communities. Their study showed that linguistic, psycholinguistic, emotional, and sentiment-based features can be useful for detecting supportive content. More recently, Kolesnikova et al. [6] explored advanced machine learning techniques for social support detection, including transformer-based models and zero-shot learning with GPT-based systems. Their results suggest that transformer models can improve the classification of supportive content. Other studies have also explored prosocial behavior in online conversations. Bao et al. [7] proposed computational methods to quantify and predict prosocial outcomes in online discussions [8] . This perspective supports the need for NLP approaches that identify constructive forms of communication. In this context, the present study contributes to the growing body of research on positive social media analysis.

This research explores the use of transformer-based models for the automatic identification of supportive comments. Transformers are neural network architectures designed to process sequential data by modeling relationships among elements in an input sequence [9].

Unlike traditional recurrent architectures, which process tokens sequentially, transformers rely on attention mechanisms to capture contextual dependencies between words, regardless of their position in the text. This allows the model to assign different levels of importance to each token according to its relevance within the sentence or document. Transformers have become particularly useful in NLP tasks because they can represent words and expressions according to their surrounding context. Therefore, transformer-based models are suitable for identifying social support, since supportive messages may be expressed through different ideas rather than through a fixed set of explicit keywords.

This proposal aims to solve three tasks for SSD-2026 [10] at the IberLEF 2026:

- **Binary Support Detection (Spanish or English)**
- **Binary Target Identification (Spanish or English)**
- **Targeted Group Multiclass Classification (Spanish or English)**

2. Data description

The dataset was constructed from comments collected from YouTube videos. The selected videos were chosen because they were likely to elicit comments expressing different forms of social support. It consists of both English and Spanish language comments.

The social support detection task was organized as a three-step classification process:

1. Identify whether a comment expresses social support.
2. Classify supportive comments according to whether the support is directed toward an individual or a group.
3. For comments expressing support toward a group, identify the specific group being supported.

For the first subtask, a comment was considered supportive if it contained a statement or message expressing encouragement, admiration, advocacy, promotion, assistance, or defense. Conversely, non-supportive comments were defined as those that did not convey any form of support according to this definition.

The second subtask determined whether the comment supported a specific person or a collective group. The classes were individual and group.

For the third subtask, group support was classified into the following categories:

- Nation: This category includes comments expressing support for a specific country, its people, or its sovereignty.
- Black community: This category includes comments that acknowledge or defend the Black community, often in relation to racial equality, social justice, or the recognition of Black cultural contributions.
- LGBTQ: This category includes comments that support the LGBTQ community by promoting equality, celebrating diversity, defending their rights, or emphasizing the need for representation.
- Religion: This category includes comments expressing support for a specific religion, religious communities, or the general right to religious freedom.
- Women: This category includes comments that highlight gender equality, recognize women's achievements, or support feminist causes.
- Other: This category includes expressions of support directed toward communities not explicitly covered by the previous categories, such as professional, activist, or other social groups.

Inter-annotator agreement (IAA) was computed to assess the consistency of the annotation process while accounting for chance agreement. Table 1 summarizes the distribution of samples across the three subtasks, where an imbalance between classes can be noticed. The imbalance ratio (IR) was 3.45, 4.20, and 52.40 for the first, second, and third English tasks, respectively. For the Spanish samples, the IR was 3.40, 3.61, and 9.33 for the first, second, and third Spanish tasks. A high IR results in a greater extent of imbalance in the dataset, making it harder to classify them [11].

Table 1
Data distribution across the proposed subtasks.

Task	Class	English samples	Spanish samples
1	Social Support	1794	590
	Non Support	6202	2010
2	Individual	345	134
	Group	1450	485
	LGBTQ	123	224
	Black Community	91	24
	Women	19	48
	Religion	15	60
	Other	416	82

Some English comment examples are shown in Table 2; the first and second rows are support clauses that correspond to group comments but in different contexts: religion and nation. The third and fourth rows are examples of non-support labels. Table 3 shows some examples from the Spanish dataset, including their translation. The first two rows correspond to a Support comment, while the rest of the rows are non-support clauses.

Table 2
English comments.

	Text	Task 1	Task 2	Task 3
1	love islam love palisten pakistan	Support	Group	Religion
2	may allah help gazans indeed allah enough	Support	Group	Nation
3	understand homophobia life people happy	Non Supp.	No	No
4	good day less messed kids single mothers	Non Supp.	No	No

Table 3
Spanish comments.

	Text	Task 1	Task 2	Task 3
1	Yo soy no binario y estoy orgulloso de ello/ I am non-binary and proud	Support	Group	LGBTQ
2	Bien por las niñas arriva México/ Good for the girls, go Mexico	Support	Group	Women
3	Solo existen dos generos../ There are only two genders..	Non Supp.	No	No
4	No entiendo/ I don't understand	Non Supp.	No	No

3. Methodology

This section outlines the specifics of the approach taken with different transformer models. A data cleaning phase was evaluated, including emoji removal, stopwords filtering, and lemmatization. However, these preprocessing steps led to a consistent decrease in performance across models. This behavior can be explained by the nature of the comments, where emojis, function words, and lexical variations often convey important emotional and contextual information. Therefore, all subsequent experiments were conducted using the original, unprocessed dataset. Figure 1 shows the proposed methodology.

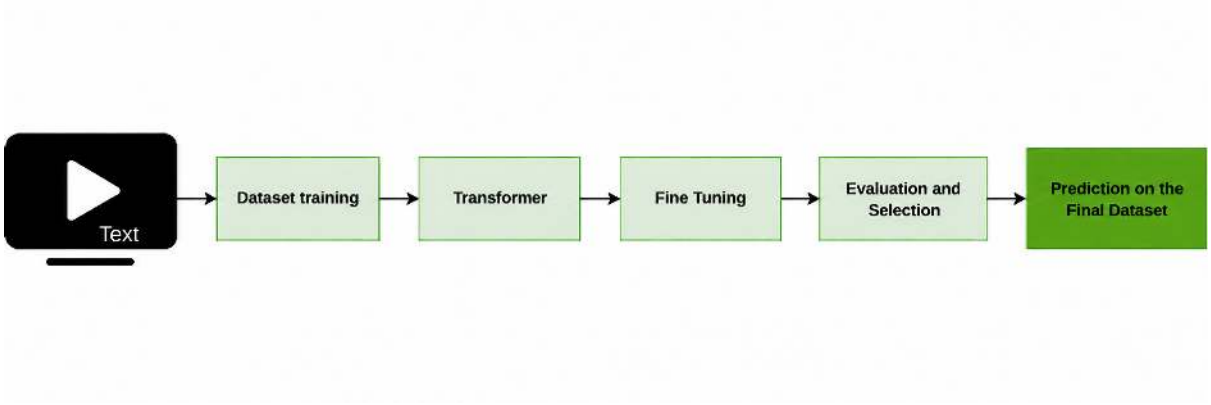

Figure 1: Methodology overview diagram

3.1. Transformer Models

The transformer network is a novel architecture that has been used in natural language processing. The model has two principal parts: the encoder and the decoder, with softmax normalization applied to them. Inputs are embedded with positional encodings to capture contextual meaning. Contextual meaning is processed by multi-head attention mechanisms and feed-forward networks in parallel to capture relationships between inputs. The decoder uses masked multi-head attention that incorporates encoder-derived attention vectors, subsequently applying feed-forward and linear layers to output probabilities [12]. One of the challenges of working with this type of model is the amount of data for training; however, strategies like fine-tuning allow unfreezing some layers of the pre-trained models and updating them with new data [13].

Diverse transformer models from Hugging Face [1] were used to solve the tasks. Table 4 shows the models employed in this study and their descriptors. BERT is a pretrained model on English data in a self-supervised fashion [14]. ALBERT is a small, parameter-efficient version of BERT model that was pretrained with a big corpus from BooksCorpus and Wikipedia [15]. DistilBERT is an uncased transformer model that was trained using BERT as a teacher through knowledge distillation [16]. BETO is a BERT model trained on a big Spanish corpus [17]. RoBERTuito is a pre-trained model trained with 500 million tweets in Spanish [18]. Every model was trained for 5 epochs, with a batch size equal to 8.

[1] https://huggingface.co/

Table 4
Transformer Models

Descriptor	Model
BERT	bert-base-uncased
ALBERT	albert-base-v2
DistilBERT	distilbert-base-uncased
BETO	dccuchile/bert-base-español-wwm-uncased
RoBERTuito	robertuito-base-uncased

4. Results

To report the best model performance on the test set, we considered the following metrics: precision, recall, accuracy, and F1-score. The values of these metrics are shown in Table 5. For Task 1, the best results for classifying the English test dataset were obtained by the DistilBERT and BERT models, with an F1 score of 0.80. For Task 2, the best performance was for BERT, and for Task 3, BERT obtained 0.86 of F1 score. Being BERT, the model that showed more consistency on the comparative study of the 3 models for classification in English, see confusion matrices from Tables 6 to 8.

Meanwhile, it was observed that BETO's performance on the Spanish dataset classification was better than RoBERTuito's across 2 of the 3 tasks. Therefore, the confusion matrix for this model is shown from Table 10 to 12.

Table 5
Models' classification of datasets under different tasks (English).

Model	Task	Precision	Recall	F1	Accuracy	Support
BERT	1	0.79	0.81	0.80	0.87	2000
	2	0.85	0.83	0.84	0.91	441
	3	0.87	0.86	0.86	0.87	363
ALBERT	1	0.80	0.78	0.79	0.86	2000
	2	0.80	0.81	0.80	0.88	441
	3	0.87	0.81	0.83	0.87	363
DistilBERT	1	0.81	0.78	0.80	0.87	2000
	2	0.85	0.79	0.82	0.90	441
	3	0.52	0.54	0.52	0.78	363

Table 6
Confusion matrix DistilBERT Task 1.

	Non-Supportive	Supportive
Non-Supportive	0.93	0.07
Supportive	0.37	0.63

Table 7
Confusion matrix BERT Task 2.

	Group	Individual
Group	0.95	0.05
Individual	0.22	0.78

Table 8
Confusion matrix BERT Task 3.

	Black C.	LGBTQ	Nation	Other	Religion	Women
Black Community	1	0	0	0	0	0
LGBTQ	0	0.94	0	0.03	0	0.03
Nation	0	0.01	0.89	0.1	0	0
Other	0.05	0.02	0.16	0.77	0	0
Religion	0	0	0.25	0	0.75	0
Women	0	0.2	0	0	0	0.8

Table 9
Models's classification of datasets under different tasks (Spanish).

Model	Task	Precision	Recall	F1	Accuracy	Support
BETO	1	0.50	0.61	0.53	0.91	651
	2	0.85	0.83	0.84	0.89	157
	3	0.55	0.86	0.57	0.89	121
RoBERTuito	1	0.83	0.81	0.82	0.88	651
	2	0.68	0.75	0.67	0.68	157
	3	0.46	0.37	0.38	0.63	121

Table 10
Confusion matrix BETO Task 1.

	Non-Supportive	Supportive
Non Support	0.96	0.04
Support	0.26	0.74

Table 11
Confusion matrix BETO Task 2.

	Group	Individual
Group	0.94	0.06
Individual	0.28	0.72

5. Discussion

The best performing models in this study were based on BERT and BETO architectures, for English and Spanish, respectively. Each model was fine-tuned for a specific target task. It was expected that RoBERTuito could have a better performance than BETO, because RoBERTuito is a model adapted for tweets' sentimental analysis, while BETO is a general model (both for Spanish); the empirical test showed something different. BETO achieved the highest performance on both training and testing sets, maybe due to the characteristics of the dataset: 650 sentences of training for task 1, 157 for task 2, and only 121 sentences for task 3. Besides the longitude of the datasets, there are other issues to consider; the IR was too high in both datasets. In English task 3, this reached 52.40, and for the Spanish samples, it reached 9.33. However, according to the final results, our work achieved 7th, 18th, and 8th positions for Spanish Tasks 1, 2, and 3, respectively. Meanwhile, the English task results were ranked 21st, 2nd, and 5th for Tasks 1, 2, and 3.

Table 12
Confusion matrix BETO Task 3.

	Black C.	LGBTQ	Nation	Other	Religion	Women
Black Community	1	0	0	0	0	0
LGBTQ	0	0.96	0.02	0	0.02	0
Nation	0.08	0.17	0.67	0.08	0	0
Other	0	0	0.05	0.95	0	0
Religion	0	0	0	0.2	0.8	0
Women	0	0	0.17	0.08	0	0.75

6. Conclusion

This study contributes to the development and evaluation of NLP systems that detect social support in social media comments and identify the target of that support (individual vs. community, and the specific community), and move beyond to an understanding of the factors that affect people in real scenarios. Due to the widespread dissemination of messages of no support through social media, it is worth developing tools to identify their impact.

The results highlight the importance of task-specific training, particularly when the source and target tasks share semantic and contextual similarities. The best F1-score results for each English task were 0.80, 0.84, and 0.87 for Tasks 1, 2, and 3, respectively; the best F1-scores for each Spanish task were 0.82, 0.84, and 0.57 for Tasks 1, 2, and 3, respectively.

7. Acknowledgments

This work was supported by Instituto Politécnico Nacional (IPN) through Secretaría de Investigación y Posgrado (SIP-IPN) research grants SIP-20260496, SIP-20260772, IND-20260569, and SIP-20260296; Comisión de Operación y Fomento de Actividades Académicas del IPN (IPN-COFAA) and Programa de Estímulos al Desempeño de los Investigadores (IPN-EDI) and Secretaría de Ciencia, Humanidades, Tecnología e Innovación, Sistema Nacional de Investigadores (SECIHTI-SNII).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.3 in order to improve the grammar and spelling of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final context and language of the publication.

References

[1] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).

[2] A. Bonet-Jover, J. 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.

[3] M. Shahiki Tash, L. Ramos, A. Zahra, A. Y. Barrera Animasas, 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] 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.

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

[7] J. Bao, et al., Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations, in: Proceedings of the Web Conference 2021, 2021.

[8] M. S. Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, G. Sidorov, et al., Online social support detection in spanish social media texts, arXiv preprint arXiv:2502.09640 (2025).

[9] N. Patwardhan, S. Marrone, C. Sansone, Transformers in the real world: A survey on nlp applications, Information 14 (2023) 242.

[10] M. Shahiki Tash, Z. Ahani, L. Ramos, A. Y. Barrera Animasas, F. Balouchzahi, O. Kolesnikova, G. Sidorov, A. Gelbukh, R. Monroy, A. Agrawal, Ssd-2026: Social support detection and target identification in social media, Shared Task at IberLEF 2026, 2026. URL: https://sites.google.com/view/ssd2026/home, official task website.

[11] R. Zhu, Y. Guo, J.-H. Xue, Adjusting the imbalance ratio by the dimensionality of imbalanced data, Pattern Recognition Letters 133 (2020) 217–223.

[12] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI open 3 (2022) 111–132.

[13] E. Kotei, R. Thirunavukarasu, A systematic review of transformer-based pre-trained language models through self-supervised learning, Information 14 (2023) 187.

[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2020).

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

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

[18] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.