International Journal of Combinatorial Optimization Problems and Informatics, 17(1), Jan-April 2026, 67-75. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i1.1239

www.editada.org

Detecting Hope in Social Media Discourse Using Machine and Deep Learning Classifiers

Ahmad Imam Amjad [1], Hamza Imam Amjad [2], Grigori Sidorov [3]*

[1] Department of Computer Science, The University of Punjab, Pakistan
[2] Department of Education, Punjab, Pakistan.
[3] Center for Computing Research, National Polytechnic Institute, Mexico
ahmadimamamjad@gmail.com, sidorov@cic.ipn.mx, hamzaimamamjad786@gmail.com

**Corresponding author.*

Abstract. Hope speech refers to messages that convey optimism, support, or expectations of a better future. With the growing use of social media for self-expression, analyzing such messages can provide meaningful insights into emotional well-being of people. However, hope speech detection has received limited attention in social media discourse analysis compared to tasks such as hate speech detection. This study addresses this gap by conducting both binary and multiclass classification of hope speech in two languages: (i) English, and (ii) Spanish. In the binary task, our goal is to distinguish between hopeful and non-hopeful tweets, while the multiclass task further categorizes hopeful content into five types: (i) *Generalized Hope*, (ii) *Realistic Hope*, (iii) *Unrealistic Hope*, (iv) *Sarcasm*, and (v) *No Hope*. We evaluate six traditional machine learning algorithms and three deep learning and transformer-based architectures. Our experimental results show that transformer models outperform traditional approaches in both languages. In English, RoBERTa achieved the highest performance (binary: 82.25% weighted F1; multiclass: 72.49% weighted F1), while in Spanish, XLM-RoBERTa showed best performance (binary: 86.32% weighted F1; multiclass: 77.01% weighted F1). These findings highlight the effectiveness of transformer-based models for multilingual hope speech detection.

Keywords: natural language processing, hope speech detection, sentiment analysis, machine learning, deep learning

Article Info
Received August 11, 2025
Accepted Oct 22, 2025

1 Introduction

Emotional expression is one of the primary aspects of human communication and one of the essential elements of psychological health. Hope is one of the existing affective states that has the power to provide people with the opportunity to visualize the future directions, spread optimism, and draw positive ways to reach the desired results. Human capacity to imagine future events has a deep impact in terms of feelings, actions, and the general mood (Bruininks & Malle, 2005). With the modern digital environment, online platforms have transformed the ways in which people interact with each other, engage in conversations, and share their experiences. In the digital age, social media now serves as a primary source of communication where users post their experiences, and spread messages that promote hope, resilience, and solidarity, and such posts are often know as the “Hope Speech”. Moreover, these open-source posts encourage kindness, unity, and positivity in people, particularly during a global crisis, such as the COVID-19 pandemic. Empirical evidence suggests that high hope levels are linked to more positive attitude, such as high academic achievement (Snyder, 2002), reduce the possibility of depressive symptoms (Snyder et al., 1997). On the other hand, individuals with low levels of hope are inversely correlated with low well-being (Diener, 2009). Consequently, it is important to understand, and analyze social media data to develop interventions to promote the well-being of people.

The fast development of online communication enhances the effectiveness of hope speech detection in mental-health analytics and discourse analysis. Automatic techniques can be very helpful in real time analysis of publicly available data on social media platforms, which in turn can provide subtle insights into the ways communities express optimism and help make more data-driven and informed decisions (Ahmad et al., 2024) and help develop more positive virtual surroundings (Aggarwal et al., 2023). Nevertheless, real time analysis of social media data using automatic techniques is not an easy task because hope is an abstract and context-dependent concept, and it may vary among cultures. There are various forms in which hope can be expressed, and this adds more complexity in the task of detecting hope speech in text. For example, O'Hara (2021) defines hope into three main categories: (i) generalized hope, that is a positive outlook without a particular goal; (ii) particularized hope, that is a goal-focused disposition based on self-efficacy; and (iii) transformative hope is a reflective construct that fosters self-knowledge and personal development. Such taxonomical differences make computational modeling more difficult, since some rhetoric often overlaps with encouragement or gratitude, or even with sarcasm. Riloff et al. (2013) showed that sarcasm often makes use of positive lexical items to derive criticism, which makes even harder the development of sentiment-analysis systems. Furthermore, users who are multilingual often tend to switch languages when they post or use code-switching in the same sentence. These types of communication can encode several different emotional states and make hope speech detection even more challenging. Therefore, it would be necessary to differentiate the subtle differences where hope is expressed in positive or ironic text, which would require context-sensitive algorithms that can handle the linguistic expressions with subtlety when dealing with multilingual environments.

To a large extent, the required machine-learning (ML) and deep-learning (DL) approaches can be explained by the fact that manual annotation of hopeful speech on social-media platforms is both labor-intensive and time-consuming. Traditional ML methods, e.g., Support Vector Machines (SVM), Logistic Regression (LR), and Random Forests, use hand-constructed textual features such as term frequency-inverse document frequency (TFIDF) or bag-of-words. Although these features can help machine learning models to provide accurate predictions to some extent, these techniques overlook the small significances and practical indications of hope as hope can be expressed in many ways. To contextually understand hope, distributed language representations can be learned through neural networks architectures, such as Convolutional Neural Networks (CNNs) and Long Short-term memory (LSTM) networks. Transformer-based models like BERT, RoBERTa, and XLM-R can enhance the correctness of emotion recognition by identifying the relationships between context and extracting the long-range relationships of words in the text. These models are particularly effective at identifying complex affective cues such as hope. Furthermore, most current studies focused on monolingual and binary tasks (hope vs. non-hope), which restrict the use of those models in other settings, such as multilingual and multicultural society of the social-media discourse (Chakravarthi and Muralidaran, 2021; Khanna et al., 2022).

To address these weaknesses, this paper examines the hope speech classification by using social-media posts in two languages: (i) English, and (ii) Spanish. We use both traditional machine learning techniques as well as deep learning neural networks classifiers. For binary and multi-class classification task, we use with multilingual annotated data of HOPE@IberLEF (Butt et al., 2025). There are five categories in the multi-class taxonomy: (i) Generalized Hope, (ii) Realistic Hope, (iii) Unrealistic Hope, (iv) Sarcasm, and (v) Not Hope. We investigate how different model architectures (e.g., monolingual vs. multilingual, transformer vs. traditional machine learning) perform on both tasks and examine their ability to generalize across languages.

2 Literature Review

The construct of hope has psychological and linguistic aspects that determine computational modeling. Bruininks and Malle (2005) define hope as an emotion that is directed towards uncertain but desirable things in the future, where there is a balance between cognitive anticipation and emotional regulation. The Hope Theory by Snyder (2002) puts hope in terms of agency, which is the desire to achieve something, and pathways, which are the perceived capacity to find ways to achieve the desired things. Academic success and good health are expected by high hope (Snyder et al., 1997), whereas low hope is associated with depression and low life satisfaction (Diener, 2009). O'Hara (2021) adds to this model by distinguishing between generalized, particularized, and transformative hope and lays stress on its many-sidedness. Cues in the hopeful discourse would include believe, will, and together, which are present and express hope and unity but differ within the culture. Such bases are critical to the modeling of the encoding of motivational and emotional states in multilingual social-media communication.

Computational research for hope speech detection was first introduced as an alternative to hate speech detection. For example, YouTube comments were analyzed for the task of hope speech detection, and the comments were relevant to the India-Pakistan conflict (Palakodety et al., 2019), who defined hope speech is a type of speech that promotes peace and respect towards each other. This work was later extended by Chakravarthi et al. (2020) to the HopeEDI corpus (English, Tamil, Malayalam), which facilitates the multilingual aspect in the datasets. In addition, LT-EDI 2021 Shared Task (Chakravarthi & Muralidaran, 2021) reported various techniques based on machine learning techniques as well as neural networks for hope speech in a binary text classification form (hope/non-hope) to encourage further research in this domain on more detailed taxonomies and multilingual corpora.

The recent developments conceptualize the hope speech as a multiclass phenomenon, which reflects the variation of the tone and intensity. For example, Balouchzahi et al. (2023) proposed PolyHope in which Generalized, Realistic, and Unrealistic Hope texts are identified in comparison with Non-Hope texts. Additionally, HOPE@IberLEF 2023 (Jiménez-Zafra et al., 2023) and HOPE@IberLEF 2025 (Butt et al., 2025) also proposed new datasets to support research in hope speech in multilingual settings (in both English and Spanish lanaguges) for both binary and multiclass hope speech text classification. These corpora present the linguistic and cultural diversity of hopeful discourse, but more data is required to make models more interpretabile (Riloff et al., 2013).

Hope speech detection has been significantly enhanced by methodological evolution. Traditional machine learning algorithms, such as logistic regression, naive Bayes, and SVM, combined with various text representation techniques such as TF-IDF, n-grams and the bag-of-words offered good performance for hope speech detection, and can be used as baselines (Chakravarthi, 2022). However, these traditional features techniques often ignore the context and semantics of text. On the other hand, neural networks-based architectures can capture more complex patterns and semantics in text. Moreover, deep learning models based on transformers (BERT, RoBERTa, mBERT, XLM-R) redefined the state of the art for the hope speech detection with the use of contextual embeddings. Mutual evaluations in LT at LT-EDI and IberLEF were able to establish that transformers score the highest in macro-F1 and can work with multilingual data effectively (Khanna et al., 2022; Das et al., 2023). Despite these benefits, cross-domain generalization of these models still requires some improvement, and it would be desirable to develop culturally adaptive and interpretable methods.

Emotional analysis across the linguistic borders has become very important currently due to the proliferation of the online communities between several languages. The Spanish speaking sources, such as Hope-Speech LGBT (Garcia-Baena et al., 2023) show that there are grammatical and pragmatic factors, such as subjunctive mood and collective framing, that have impact on the expression of hope. Moreover, multilingual transformers-based models are not as good as monolingual models (e.g., BETO), but different prompting techniques can be used such as zero-shot, which provide better results (Sidorov et al., 2025). Furthermore, Hybrid systems (i.e., deep embeddings + psycholinguistic features such as LIWC, NRC lexicons) can be used to enhance the explanatory power and cultural rooting of deeply embedded systems (Arif et al., 2025). Active learning, weighting classes, explainable-AI (e.g., LIME) are other techniques that increase transparency and balance in the classes. Thus, hope speech detection in bilingual setting (English and Spanish language comparisons) can serve as a significant step in an inclusive NLP that is able to identify subtle culturally specific manifestations of hope.

3 Methodology

The purpose of this work is to develop and evaluate automatic techniques that can detect hope speech in social media text in two languages: (i) English, and (ii) Spanish. In this proposed methodology, we use a combination of traditional machine learning and deep learning for two tasks. In the binary settings, the task is to distinguish between hopeful and non-hopeful tweets, while the in multiclass settings, the task is to further categorizes hopeful content into five types: (i) *Generalized Hope*, (ii) *Realistic Hope*, (iii) *Unrealistic Hope*, (iv) *Sarcasm*, and (v) *No Hope*. Figure 1 shows the implementation and evaluation process of machine and deep learning techniques for hope speech detection.

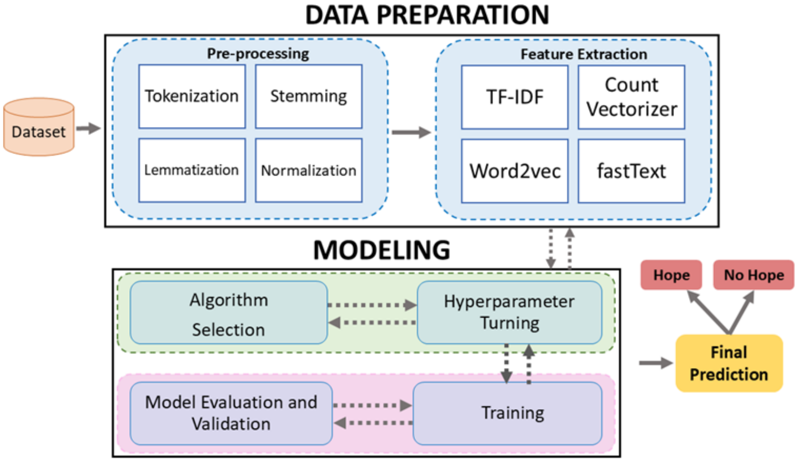

Figure 1. The implementation process of machine and deep learning modeling.

3.1 Dataset

The dataset contains tweets in two languages (English and Spanish). The multilingual dataset proposed by Butt et al. (2025a) contained tweets. Each tweet is annotated with one of five labels: (i) *Generalized Hope*, (ii) *Realistic Hope*, (iii) *Unrealistic Hope*, (iv) *Sarcasm*, and (v) *No Hope.* In this dataset, a new category of hope "sarcasm" was also included, which is very helpful to differentiate subtle differences of real hope from ironic or mocking expressions.

The English subset was derived from the original PolyHope dataset (Balouczhahi et al., 2022) that contained tweets collected from Twitter between January and June 2022. The dataset supports two tasks: (a) binary classification, where each post is labelled as Hope or Not Hope, and (b) multiclass classification, with five categories: Generalized, Realistic, Unrealistic, Sarcasm, and Not Hope. The dataset is not balanced. The majority of the posts are called Not Hope, and such labels as Sarcasm or Unrealistic Hope are less frequent. PolyHope V2 is open-source on the shared task platform and is widely used as a general multilingual emotion recognition model, especially in separating fine affective change like irony and varied versions of hope (Balouczhahi et al., 2025).

Table 1. Tweets distribution in the datasets (Butt et al., 2025a; 2025b).

Task	Label	English Tweets	Spanish Tweets
Binary	Hope	4,434	9,654
	Not Hope	5,081	10,788
Multiclass	Not Hope	4,081	9.788
	Sarcasm	1,259	1,259
	Realistic Hope	982	2,024
	Generalized Hope	2,335	5,007
	Unrealistic Hope	858	2,364
Total		9,515	20,442

3.2 Data Preprocessing

The preprocessing of data differed depending on the kind of model. The posts are usually informal, long, and include emojis, hashtags, mentions of users, and other social-media artifacts. Cleaning was performed on a basic level: URLs, mentions, and special characters were deleted, and emojis were turned into text. To convert the text to lower case, stop words, HTML tags, usernames, URLs, numbers, hexadecimal regular expressions, punctuation marks were removed, extra spaces were collapsed, the text was tokenised, all stop words and non-alphabetic tokens were deleted, words were stemmed, and tokens combined into one string, which was processed as a traditional machine learning model. There was also the usage of emojis in written form.

The text was then tokenized and lemmatized using the NLTK and Spacy libraries. After these steps, the text was turned into a numerical representation using TF-IDF. To keep the data manageable, only the top 5,000 to 10,000 most common words were used in the training vocabulary. Moreover, the text was kept mostly in its original form for deep learning models. This is because these models can better understand context of words when the structure of the text is preserved. Punctuation, capital letters, and word order were all kept in the text. Tokenization was done using the model's own tokenizer. The text was either cut or inflated to a standard number, typically 128 or 512 tokens. We distributed the data into two parts: we used 70% of the original data for the training purpose and 30% of the original data was used for the validation and testing.

3.3 Feature Representation

Different features representations techniques were employed in this study. In the case of traditional machine learning models, the Sklearn library along with Python programming language were used. To represent words into vectors, we used three text vectorization techniques: (i) CountVectorizer (generating a Bag-of-Words frequency matrix), (ii) TF-IDF (Term Frequency-Inverse Document Frequency) at word-level, character-level as well as unigrams and bigrams, and (iii) HashingVectorizer to transform text into a numerical form. For the implementation of deep learning models, the Gensim library was used. We used two word embedding models to generate word embeddings: (i) Word2Vec (capturing context-based meaning), and (ii) FastText (using sub-word embeddings to carry more meaning and context). In some cases, extra features were added to the transformer-based models. These included sentiment scores, future-related words such as "will" or "might" the types of emojis used, and grammar patterns such as parts of speech or named entities.

3.4 Classification Models

We used both machine learning and deep learning techniques for hope speech text classification in English and Spanish language and to evaluate the best-performing techniques for both binary classification (hope vs. not hope) and multiclass classification (generalized hope, realistic hope, unrealistic hope, and sarcasm). For traditional machine learning, we used six machine learning models: (i) Naive Bayes, (ii) Random Forest, (iii) Logistic Regression, (iv) Linear SVM, (v) Decision Tree, and (vi) Gradient Boosting. On the other hand, we used three neural network-based and transformer-based models: (i) DistilBERT, (ii) RoBERTa, and (iii) XLMRoBERTa. These models were fine-tuned for both binary and multiclass classification tasks. For binary tasks, the final layer was configured to produce two outputs (hope, not hope).

For multiclass text classification, the output layer was adjusted to produce five outputs corresponding to different types of hope and sarcasm for multiclass tasks. The training of these models was performed using the AdamW optimizer, with learning rates between 2e-5 and 5e-5. Training was conducted in small batches, and early stopping was employed to prevent overfitting. To further enhance the results, some transformer-based models combined the output with additional features to improve classification performance on more challenging examples, which show better results on harder-to-classify instances. For models training and optimization and to handle the problem of class imbalance in the dataset, we used stratified sampling, class-weighted loss functions, and special loss methods like focal loss. Training was done over three to four rounds, and performance was measured after each round. We used grid search for hyperparameters and three-fold cross-validation in our experiments settings.

3.5 Models Evaluation

To check the robustness of the models in real-life settings and examine their ability to make predictions on unseen data, we used several evaluation metrics. (i) precision, (ii) recall, (iii) accuracy, (iv) macro F1-score, and (v) weighted F1-score. We used these metrics to view the overall performance of different models as well as their ability to handle class imbalance. Additionally, we also calculated precision and recall scores separately for each class to evaluate the ability of models to correctly detect positive instances and ignore false negatives for both tasks in both binary and multiclass settings.

Furthermore, confusion matrices were used to visualize where the models made incorrect predictions, and to highlight areas of improvement such as false positives and false negatives. This evaluation and analysis helped in understanding whether certain classes, such as Sarcasm or Unrealistic Hope, were more difficult to classify and whether the models were prone to misclassifications in those categories. For the binary classification task (hope vs. not hope), ROC-AUC scores were also calculated to evaluate the models' ability to separate the two classes effectively.

3.6 Results

The results of binary and multiclass hope speech classification tasks are presented for each class separately for both English and Spanish languages (see Tables 1, 2, 3, and 4). Several machine-learning and transformer-based models were tested and evaluated for the English and Spanish languages. The task aimed to identify whether a social media post expresses hope and to classify the type of hope in the multiclass settings. Special focus was given to difficult classes like “Sarcasm” and “Unrealistic Hope”, which are often difficult to detect in hope speech text classification task.

3.6.1 Performance of Models on Binary Hope Speech Classification

In the binary classification task, the goal was to distinguish between “hope” and “not hope”. We used six machine learning models, and among all the models, Random Forest achieved the best results for the English language and obtained an accuracy of 80.42% accuracy and 80.41% weighted F1score for the binary hope speech task. Other traditional machine learning models such as Logistic Regression were faster to train and easier to interpret, which made them useful in scenarios where speed and simplicity were more important than achieving the highest possible accuracy.

Other models such as Logistic Regression showed poor performance (achieved 77.76% accuracy and weighed F1score of 77.75% in the binary task), and linear SVM showed poor results (achieved 72.89% accuracy and weighed F1score of 72.89% in the binary task) compared to all other machine learning models. For Spanish language, Random Forest model also achieved the highest performance, with an accuracy of 85.93% and weighed F1score of 85.93%. These models are beneficial in real-time applications or scenarios where interpretability and training speed are crucial.

The experiments results revealed that the transformer-based models outperformed all traditional techniques in both languages and for both tasks (binary and multiclass). In deep learning, RoBERTa showed the best performance compared to other neural networks models and achieved accuracy of 82.24% and weighed F1score of 82.25%. RoBERTa is a transformer-based model and a variant of BERT (Bidirectional Encoder Representations from Transformers). It was designed to remove certain training constraints that BERT experience as well as it should be able to be trained on a big dataset so that a better performance of model can be achieved. RoBERTa uses bidirectional attention mechanisms, which means that when words are encoded, it takes into account both left and right contexts.

On the other hand, traditional machine learning classifiers only depend on a sequential (unidirectional) view of the text. As a result of deep contextual understanding, RoBERTa understand the meaning of a word or phrase based on the entire context in which it occurs. Furthermore, RoBERTa's transformer layers consist of multiple attention heads that capture complex, high-level interactions between words. For example, based on pattern availabke in the training data, RoBERTa has the ability to identify when sarcasm or humor is involved in the text because it has the ability to use bidirectional attention mechanisms.

RoBERTa's self-attention mechanism allows it to capture relationships between words even when those relationships are spatially distant in a sentence. Traditional models such as Naive Bayes or Random Forest do not have this ability because they treat each feature independently. For example, in sarcasm detection, RoBERTa can distinguish between literal and sarcastic hope based on contextual shifts in tone and structure, which other models may miss. In contrast, traditional models like Logistic Regression or SVM rely on a simplified bag-of-words approach that does not capture the full context of each word. Similarly, on the Spanish dataset, XLM-RoBERTa achieved an accuracy of 86.32% and weighted F1-score of 0.86.32%, while DistilBERT achieved 81.65% accuracy and weighted F1-score of 81.66%. Table 2 and 3 show the performance of both machine and deep learning models across both English and Spanish languages for the binary Classification task.

Table 2. Binary Classification Performance across Models in English language.

Model	Precision	Recall	Accuracy	Macro F1	Weighted F1
Machine Learning Models					
Naive Bayes	0.7676	0.7686	0.7685	0.7679	0.7687
Random Forest	0.8034	0.8028	0.8042	0.8031	0.8041
Logistic Regression	0.7766	0.7763	0.7776	0.7764	0.7775
Linear SVM	0.7276	0.7275	0.7289	0.7276	0.7289
Decision Tree	0.7363	0.7349	0.7373	0.7354	0.7369
Gradient Boosting	0.8043	0.7981	0.8021	0.7996	0.8011
Deeps Learning Models					
DistilBERT	0.8156	0.8167	0.8165	0.8159	0.8166
RoBERTa	0.8215	0.8225	0.8224	0.8219	0.8225
XLM-RoBERTa	0.8127	0.8133	0.8137	0.8130	0.8137

Table 3. Binary Classification Performance across Models in Spanish language.

Model	Precision	Recall	Accuracy	Macro F1	Weighted F1
Machine Learning Models					
Naive Bayes	0.7981	0.7996	0.7981	0.7979	0.7983
Random Forest	0.8590	0.8586	0.8593	0.8588	0.8593
Logistic Regression	0.8298	0.8303	0.8304	0.8300	0.8305
Linear SVM	0.8275	0.8279	0.8281	0.8276	0.8281
Decision Tree	0.8075	0.8081	0.8080	0.8077	0.8081
Gradient Boosting	0.7841	0.7812	0.7835	0.7819	0.7830
Deeps Learning Models					
DistilBERT	0.8282	0.8288	0.8268	0.8268	0.8279
RoBERTa	0.8278	0.8271	0.8240	0.8240	0.8239
XLM-RoBERTa	0.8658	0.8658	0.8632	0.8632	0.8632

3.6.2 Performance of Models on Multiclass Hope Speech Classification

In the multiclass classification task, the goal was to classify each input into one of five categories: Generalized Hope, Realistic Hope, Unrealistic Hope, Not Hope, and Sarcasm. Several machine learning models and deep learning models were trained and evaluated on both English and Spanish datasets. We used six machine learning classifiers, and most models showed moderate results. For example, Naive Bayes had the lowest performance and obtained accuracy of 54.65% and 45.90% weighed F1-score for English language as well as for Spanish language and achieved 64.38% and 58.36% weighed F1-score. In contrast, linear SVM and the Voting Classifier showed slightly better performance. Logistic Regression and Random Forest also showed moderate performance.

On the other hand, transformer-based neural networks showed better performance compared to traditional machine learning techniques. For example, DistilBERT performed well in the binary classification task but had a lower macro F1 score in the multiclass setting. For multiclass classification task, RoBERTa outperformed all other classifiers and obtained 72.10% accuracy and 72.49% weighted F1-score in English language. The transformer models showed better performance in recognizing subtle differences between the types of hope in the text and were very efficient in detecting sarcasm, which is often considered as a challenging category for both humans and machines. This shows that transformer models can capture nuanced language patterns that are often ignored by the traditional and other neural networks models.

Table 4. Multiclass Classification Performance across Models in English language.

Model	Precision	Recall	Accuracy	Macro F1	Weighted F1
Machine Learning Models					
Naive Bayes	0.8221	0.3216	0.5465	0.3027	0.4590
Random Forest	0.6624	0.5769	0.6868	0.5880	0.6613
Logistic Regression	0.6102	0.6487	0.6679	0.6244	0.6723
Linear SVM	0.6018	0.6063	0.6553	0.6039	0.6574
Decision Tree	0.5337	0.5374	0.5849	0.5348	0.5912
Gradient Boosting	0.6373	0.5703	0.6768	0.5790	0.6541
Deeps Learning Models					
DistilBERT	0.6583	0.6377	0.7026	0.6466	0.7003
RoBERTa	0.6808	0.6832	0.7210	0.6810	0.7249
XLM-RoBERTa	0.6670	0.6345	0.7126	0.6472	0.7079

Table 5. Multiclass Classification Performance across Models in Spanish language.

Model	**Precision**	**Recall**	**Accuracy**	**Macro F1**	**Weighted F1**
Machine Learning Models					
Naive Bayes	0.7406	0.4110	0.6438	0.4386	0.5836
Random Forest	0.8235	0.6622	0.7706	0.7183	0.7591
Logistic Regression	0.6474	0.7090	0.6953	0.6708	0.7033
Linear SVM	0.7078	0.7239	0.7370	0.7151	0.7391
Decision Tree	0.6768	0.6961	0.7077	0.6859	0.7104
Gradient Boosting	0.6184	0.4533	0.6266	0.4944	0.5926
Deeps Learning Models					
DistilBERT	0.6418	0.6265	0.6907	0.6334	0.6895
RoBERTa	0.6705	0.6740	0.7242	0.6704	0.7112
XLM-RoBERTa	0.7303	0.7434	0.7645	0.7348	0.7701

3.7 Error Analysis

In our experiments, we observed that both traditional machine learning and deep learning models performed well in recognizing hope. However, we identified some errors during the evaluation process that highlighted areas where the models faced difficulties in prediction. Although RoBERTa was very efficient in both the binary and multiclass tasks for both English and Spanish, it still experienced confusion between some types of hope, such as Generalized Hope, Realistic Hope, and Unrealistic Hope. One possible reason for this is that hope can be expressed in many ways, and these categories often share very similar language and keywords, such as “hope”, “wish”, and “maybe”. Therefore, the model frequently misclassified these posts and mislabeled them during the prediction stage. For example, the sentence 'I hope things will get better' is inherently hopeful, but it does not always fit neatly into one type of hope. This confusion caused the models to make incorrect predictions, particularly when the distinction between realistic and unrealistic hope was subtle.

Although transformer-based models (e.g., RoBERTa) showed good performance in recognizing some types of hope, such as sarcasm, posts containing sarcastic hope were still incorrectly predicted as “Not Hope” or “Unrealistic Hope”. Sarcasm can be expressed in different ways, and detecting it can be challenging, especially since it is often conveyed through optimistic words that carry contrary meanings. For example, the sentence “Just what I needed – another reason to hope!” appears positive, but it is an example of sarcasm. This can lead models to make incorrect predictions when classifying sarcasm. Moreover, class imbalance was another important issue. Some types of hope had a low frequency of posts in certain categories, such as Realistic Hope and Unrealistic Hope. Since these classes had very few samples, traditional machine learning techniques, such as Logistic Regression and even DistilBERT, struggled to learn these categories and make accurate predictions. This is why many posts in these categories were incorrectly classified as more generic categories, like “Not Hope” or “Generalized Hope”. This suggests that class imbalance also contributed to incorrect or misleading predictions. Furthermore, many posts in the dataset contained slang, emojis, and abbreviations, which added additional complexity to the classification task. Machine learning classifiers, such as SVM and Logistic Regression, experience issues with predictions when encountering this type of informal language. In contrast, DistilBERT was better at handling slang; however, it made numerous incorrect predictions in multiclass classification. Short and ambiguous posts also posed another issue. For example, a post like “Still waiting to be given a sign” could be interpreted as hopeful or neutral, depending on the context. Despite being state-of-the-art models, RoBERTa also struggled with such indistinct posts and misclassified them.

4 Conclusion

This paper explored the task of multilingual hope speech detection. We utilize both machine learning and deep learning models on the English and Spanish data from the HOPE@IberLEF 2025 shared task. Our experiment results showed that deep learning models, particularly transformer-based models such as RoBERTa, outperformed all other classifiers in both binary and multiclass classification tasks. This shows that transformer-based models are good at capturing complex language patterns, especially in distinguishing subtle between different types of hope and sarcasm. Future research should focus on several directions. First, cross-lingual flexibility should be enhanced by models' cross-lingual capabilities through language-specific fine-tuning, as this would enable better adaptation to diverse linguistic features across languages.

Additionally, incorporating more features such as syntactic structures, sentiment scores, and emotion detection could improve performance in more nuanced text classification tasks. Lastly, developing models that can better handle ambiguity, detect sarcasm, and interpret informal languages such as slang, abbreviations, and emojis would further enhance the robustness of these models in real-world applications.

References

Aggarwal, P., Das, A., & Chakravarthi, B. R. (2023). Multilingual hope-speech detection using transformer-based models. In *Proceedings of LT-EDI 2023.* Association for Computational Linguistics. https://aclanthology.org/2023.ltedi-1.38

Ahmad, M., Shahiki-Tash, M., Jamshidi, A., & colleagues. (2024). Analyzing hope speech from psycholinguistic and emotional perspectives. *Scientific Reports*, *14*, 23548. https://doi.org/10.1038/s41598-024-74630-y

Arif, M., Shahiki-Tash, M., Jamshidi, A., et al. (2024). *Analyzing hope speech from psycholinguistic and emotional perspectives. Scientific Reports*, *14*, 23548. https://doi.org/10.1038/s41598-024-74630-y

Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2022). Polyhope: Two-level hope speech detection from tweets, DOI: 10.48550. *arXiv Preprint*. https://arxiv.org/abs/2210.14136

Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2023). PolyHope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, *225*, Article 120078. https://doi.org/10.1016/j.eswa.2023.120078

Bruininks, P., & Malle, B. F. (2005). Distinctive features of hope and related emotions. *Cognition & Emotion*, *19*(2), 113–142. https://doi.org/10.1080/02699930441000292

Butt, S., Balouchzahi, F., Amjad, A. I., Amjad, M., Ceballos, H. G., & Jiménez-Zafra, S. M. (2025a). Optimism, Expectation, or Sarcasm? Multi-Class Hope Speech Detection in Spanish and English. *arXiv Preprint.* https://arxiv.org/abs/2504.17974

Butt, S., Balouchzahi, F., Amjad, M., Jiménez-Zafra, S. M., Ceballos, H. G., & Sidorov, G. (2025b). Overview of PolyHope at IberLEF 2025: Optimism, Expectation or Sarcasm?. *Procesamiento del Lenguaje Natural*, *75*, 461–474. https://investigacion.ujaen.es/documentos/68ee90ae3b2e3f6b4f68f3db?lang=en

Chakravarthi, B. R. (2020). *HopeEDI: A multilingual hope-speech detection dataset for equality, diversity, and inclusion.* In *Proceedings of PEOPLES 2020* (pp. 41–53). ACL. https://doi.org/10.18653/v1/2020.peoples-1.5

Chakravarthi, B. R. (2022). Hope-speech detection in YouTube comments. *Social Network Analysis and Mining*, *12*(1), 75. https://doi.org/10.1007/s13278-022-00901-z

Chakravarthi, B. R., & Muralidaran, V. (2021). *Findings of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion.* In *Proceedings of LT-EDI 2021* (pp. 61–72). ACL. https://doi.org/10.18653/v1/2021.ltedi-1.8

Diener, E. (2009). *The science of well-being: The collected works of Ed Diener.* Springer.

Fazlfrs. (2025, October 19). PolyHope at IberLEF 2025: Optimism, expectation or sarcasm?. CodaBench. https://www.codabench.org/competitions/5509/

García-Baeza, D., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., García-Díaz, J. A., & Valencia-García, R. (2023). Hope speech detection in Spanish: The LGBT case. *Language Resources and Evaluation*, *57*, 1487–1514. https://doi.org/10.1007/s10579-023-09638-3

Jiménez-Zafra, S. M., et al. (2023). *Overview of HOPE@IberLEF 2023: Multilingual hope-speech detection. Procesamiento del Lenguaje Natural, 71*, 289–300. https://doi.org/10.26342/2023-71-29

Khanna, P., Das, A., & Chakravarthi, B. R. (2022). *Transformer-based approaches for hope-speech detection.* In *Proceedings of LT-EDI 2022* (pp. 423–431). ACL. https://aclanthology.org/2022.ltedi-1.49

O'Hara, D. J. (2021). Three spheres of hope: generalised, particularised and transformative. In L. Ortiz, & D. O'Hara (Eds.), *Phoenix rising from contemporary: Global society* (pp. 3–14). Brill.

Palakodety, S., KhudaBukhsh, A. R., & Carbonell, J. G. (2019). Hope-speech detection: Helping online communities become more inclusive. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 235–243). ACM. https://doi.org/10.1145/3292522.3326032

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). *Sarcasm as contrast between a positive sentiment and negative situation. EMNLP 2013* (pp. 704–714). https://aclanthology.org/D13-1066/

Sidorov, G., Balouchzahi, F., Ramos, L., Gómez-Adorno, H., & Gelbukh, A. (2025). *Multilingual identification of nuanced dimensions of hope speech in social-media texts (MIND-HOPE). Scientific Reports*, *15*(1), 26783. https://doi.org/10.1038/s41598-025-10683-x

Snyder, C. R. (2002). Hope theory: Rainbows in the mind. *Psychological Inquiry*, *13*(4), 249–275. https://doi.org/10.1207/S15327965PLI1304_01

Snyder, C. R., Harris, C., Anderson, J. R., et al. (1997). The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology*, *60*(4), 570–585. https://doi.org/10.1037/0022-3514.60.4.570