Hope Speech Detection Using Transformers and Large Language Models: A Bilingual Approach at IberLEF 2025

Diana P. Madera-Espíndola[1,*,†], Zoe Caballero-Domínguez[1,†], Valeria J. Ramírez-Macías[2,†], Sabur Butt[2,3] and Hector G. Ceballos[2,3]

[1] *Tecnológico de Monterrey, Estado de México, México*
[2] *Tecnológico de Monterrey, Monterrey, México*
[3] *Institute for the Future of Education, Monterrey, México*

Abstract

This paper presents our approach to Binary and Multiclass Hope Speech Detection within the IberLEF 2025 shared task. The objective of this study is to assess the effectiveness of various large language models (LLMs)—including GPT (ChatGPT o3-mini), Claude (3.5 Sonnet), Llama (4), and DeepSeek(R1)—by comparing their performance against XLM-RoBERTa, which we use as a baseline multilingual transformer model. To enhance the performance of these models, we leverage advanced techniques such as one-shot and few-shot learning prompting strategies, as well as data preprocessing and augmentation methods. Our results provide empirical insights into the comparative strengths and limitations of these models in detecting hope speech and distinguishing it from sarcasm across multiple languages. In Task 1, our best model achieved F1 scores of 0.8584 for English and 0.8435 for Spanish, while in Task 2, it attained 0.7563 for English and 0.7540 for Spanish.

Keywords

Hope Speech Detection, Sarcasm, Transformers, Large Language Models, Prompt Engineering, Data Augmentation

1. Introduction

In the current digital age, social media platforms have evolved into rich sources of information that can be analyzed for various purposes, including sentiment analysis and emotion identification. These platforms provide valuable insights into public opinion, emotional expression, and social dynamics, making them key data points for understanding human behavior and interactions in the digital space.

Hope can be defined as a desire or expectation oriented toward the future, relating to a specific or general event or outcome, with a significant impact on human emotions, decisions, and behavior [1]. Detecting hope speech usually involves classifying sentences as either hope speech or non-hope speech [2][3][4], however, one of the challenges in detecting hopeful speech arise from the complexity and ambiguous of this emotion that appears in diverse ways depending on the context and involves multiple emotional layers. Identifying hope becomes even more difficult when trying to accurately distinguish genuine hope from sarcasm [5], since sarcastic remarks often use positive words to express negative feelings. This gap between literal expression and true intent makes it difficult for machine learning and deep learning models to accurately detect hope [6].

In the field of automated hope speech detection, the growing popularity of transformer-based models and large language models (LLMs) has significantly advanced text representation techniques. These models, known for their ability to capture complex linguistic patterns and contextual meaning, have expanded the potential for classification [7][8][6][9][10][11][12]. Then, the problem addressed in

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

A01025835@tec.mx (D. P. Madera-Espíndola); A01747247@tec.mx (Z. Caballero-Domínguez); A01636965@tec.mx (V. J. Ramírez-Macías); saburb@tec.mx (S. Butt); ceballos@tec.mx (H. G. Ceballos)

https://github.com/DianaPME (D. P. Madera-Espíndola); https://github.com/ZoeCD (Z. Caballero-Domínguez); https://github.com/JassiveRamirez (V. J. Ramírez-Macías); https://github.com/saburbutt (S. Butt)

0009-0007-5040-6171 (D. P. Madera-Espíndola); 0009-0003-5472-1516 (Z. Caballero-Domínguez); 0009-0005-3392-3594 (V. J. Ramírez-Macías); 0000-0002-4056-8923 (S. Butt); 0000-0002-2460-3442 (H. G. Ceballos)

this paper is the detection of hope speech in both English and Spanish texts, within the context of the PolyHope shared task: Optimism, Expectation or Sarcasm?[13] at IberLEF 2025 [14]. The task focuses on analyzing how hope is expressed in social media, recognizing that hope is a complex and fundamental human emotion. However, detecting hope in text poses significant challenges for NLP systems, especially when it is masked by sarcasm or expressed in subtle ways.

This task is particularly relevant because it pushes the boundaries of existing research in emotion detection and inclusive language technologies. Most current datasets either rely on simple binary classification (hope vs. non-hope) or fail to capture nuanced subcategories such as generalized optimism, unrealistic hope, or sarcastic hope. Additionally, few resources explicitly annotate sarcasm, which limits the ability to train models that can distinguish genuine hope from sarcastic expressions. The inclusion of Spanish alongside English further addresses the lack of cross-linguistic alignment in existing annotation schemes, enabling the development of more robust multilingual models.

Our study focuses on comparing the performance of advanced models such as ChatGPT (o3-mini), Claude (3.5 Sonnet), Llama (4), and DeepSeek (R1), to a more traditional transformer: RoBERTa. We evaluate the strengths and limitations of each model in both binary (Hope or Not Hope) and multiclass classifications (Not Hope, Generalized Hope, Realistic Hope, Unrealistic Hope, or Sarcasm) for detecting hope speech and sarcasm.

Furthermore, the paper investigates the impact of various strategies aimed at optimizing model performance. These strategies include one-shot and few-shot learning techniques to enhance the LLMs' response and data augmentation methods to improve RoBERTa's ability to learn from our data. Through the evaluation of these approaches, the study aims to provide a comprehensive analysis of how advanced models can be effectively applied to detect hope speech and sarcasm in diverse linguistic contexts.

Given the enormous amount of data LLMs are trained on, we believe they are able to detect complex emotions in text, regardless of the language. Compared to traditional transformers, LLMs are trained over multiple types of texts, varying significantly in style and tone. Therefore, we believe recent LLMs such as GPT, Claude, Llama, and DeepSeek can be better detect hope and differentiate it from sarcasm than a transformer such as RoBERTa.

The research questions that we are trying to solve are the following: (1) How effective are various large language models (LLMs) such as GPT, Claude, Llama, and DeepSeek, in detecting hope speech and sarcasm in both English and Spanish texts? How do they compare to RoBERTa, a transformer? (2) What is the impact of one-shot and few-shot learning prompting strategies on the performance of LLMs for detecting hope speech and sarcasm? (3) How does data augmentation influence the accuracy and reliability of LLMs in classifying texts as hope speech and sarcasm? (4) What are the differences in performance between binary and multiclass classification approaches for hope speech and sarcasm detection?

2. Related Work

Sentiment analysis and emotion detection are central tasks in natural language processing (NLP), with applications across various domains including social media monitoring, customer feedback analysis, and social behavior prediction [15]. Early approaches to sentiment analysis primarily involved machine learning models such as Support Vector Machines (SVMs), Logistic Regression, and Naive Bayes, which were successful in binary classification tasks like determining whether a text expressed positive or negative sentiment [16] [17] [18]. With the advent of deep learning techniques, more complex models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) began to improve the accuracy of sentiment classification by capturing deeper, hierarchical features of text [19] [20].

More recently, the emergence of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers) and its derivatives like RoBERTa, and DistilBERT, has revolutionized the field of NLP. These models utilize pre-trained contextual embeddings that capture complex syntactic

Our code is publicly available at https://github.com/DianaPME/PolyHope-IBERLEF-2025 and semantic relationships within the text. They have been shown to outperform traditional machine learning models and RNN-based approaches in a wide range of tasks, including sentiment analysis, emotion detection, and text classification [21]. Recent work has applied transformer models to specific domains, such as hate speech detection. For example, a study by Malik et al. [22] utilizes a transfer learning approach by fine-tuning the RoBERTa model. In this research, XLM-RoBERTa is used to represent the data, enabling a deeper understanding of the text's nuances across different languages, specifically English and Russian.

On the other hand, the rise of Generative AI tools, especially Large Language Models (LLMs), has had a significant impact on sentiment analysis research. LLMs not only compete with traditional transfer learning models but often outperform them in sentiment classification accuracy [23]. A study by Thuy and Thin [24] highlights the advantages of using large language models (LLMs), such as ChatGPT 3.5, by exploring several prompting techniques (zero-shot prompting, few-shot prompting, and chain of thought prompting) to optimize the models' ability to generate relevant responses. All approaches demonstrated strong performance, with only slight variations in classification outcomes across the different strategies.

Additionally, RamakrishnaIyer et al. [25] have identified the challenge of data imbalance in tasks like hope speech detection, where datasets are often dominated by Non-Hope speech. To address this issue, this approach enhances model performance through data augmentation techniques such as back-translation, which was utilized to generate synthetic data by translating sentences into another language and then translating them back into the original language. This approach helps to create a more balanced dataset, especially for the underrepresented Non-Hope class.

Regarding sarcasm, which is a particular challenge for hope speech detection due to context and irony, as it involves using words that convey the opposite of their literal meanings. Research, such as the one conducted by Zhou [26], showed that using the transformer model RoBERTa, as opposed to a traditional CNN approach, greatly enhances the effectiveness of sarcasm detection. However, comprehensive evaluations such as the one made by Zhang et al. [27] revealed that while GPT-4 excels, significant improvements are still needed for LLMs to effectively understand and detect human sarcasm.

3. Data and methods

3.1. Data

The dataset employed in this study originates from the PolyHope-M track at IberLEF 2025 . It comprises Twitter texts in both English and Spanish and was divided into three subsets: a training set, a development set, and a final test set. The development and testing datasets included three columns: one for the tweet text, one for the binary label (Hope or Not Hope), and one for the multiclass label (Generalized Hope, Realistic Hope, Unrealistic Hope, Not Hope, or Sarcasm). The English training set consisted of 5,233 samples, whereas the Spanish training set contained 11,243 samples. The development set comprised 1,902 samples for English and 4,088 for Spanish. The test datasets included only a single column with the tweet text. The English test set contained 2,380 samples, and the Spanish test set contained 5,111 samples. It is worth noting that there was an imbalance between the English and Spanish datasets, which is an important factor to consider when evaluating the model's performance across both languages.

Figures 1 and 2 present the data distribution for the English and Spanish training sets, categorized into binary and multiclass labels. In the binary classification, the distribution is relatively more balanced in both languages. However, for the multiclass classification, the "Not Hope" class is the most dominant, with "Generalized Hope" being the second most frequent, though it represents only about half the number of "Not Hope" samples. The remaining three classes are more evenly distributed but are considerably smaller in proportion across the entire dataset.

Figures 3 and 4 illustrate the most frequently occurring words in the training sets for both the English

https://www.codabench.org/competitions/5509/

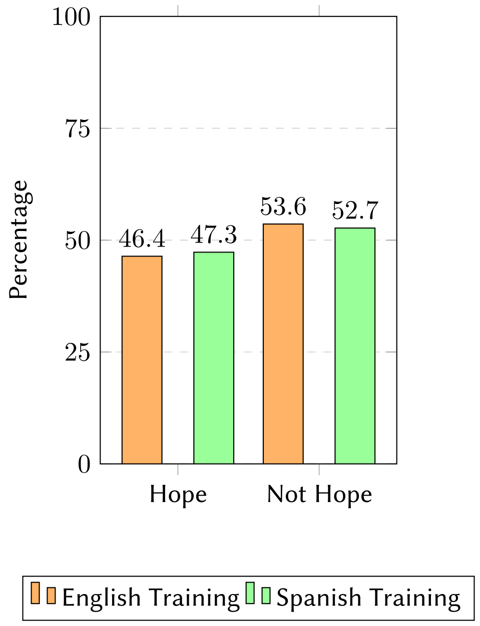


Figure 1: Binary Distribution

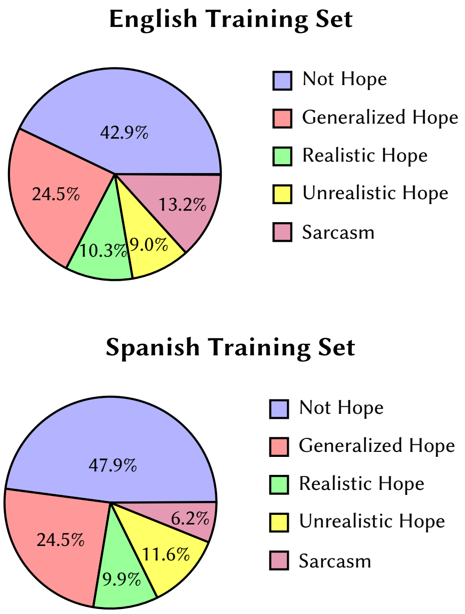


Figure 2: Multiclass Distribution

and Spanish languages. As observed, certain words like "user" are relatively common, which is expected given the nature of Twitter, where user mentions and interactions are frequent. Additionally, the abbreviation "rt" appears in the word clouds, although with less frequency, reflecting retweets and the sharing of content within the platform. Moreover, words such as "URL" and "https" are prominently featured, highlighting the prevalence of links in tweets, as well stopwords that do not have a significant meaning appear in the figures. On the other hand, terms like "hopeful," "wish," "expect," and "desired," along with their Spanish counterparts such as "esperar," "desearía," and "anhelo," are more relevant for detecting hope speech.

3.1.1. Data processing

Our first step in the methodology was to clean the data in order to enhance the performance of the models. The cleaning process involved:

- Converting text to lowercase and remove spaces: This step helps standardize the format of the text data, which is essential for consistent text processing and comparison.
- Removing HTTP links: URLs do not contribute meaningfully to the semantic understanding of the text and could add noise, so we removed them to focus on the actual content.
- Removing Twitter mentions (user) and retweet (rt): Mentions to specific users or retweets are typically context-dependent and do not contribute to the broader semantic analysis.
- Removing non-alphabetical characters: We retain only the following characters in the text: 'abcdefghijklmnñopqrstuvwxyz!?#0123456789 '. This is done to eliminate corrupted characters, such as Arabic symbols, that were detected during preprocessing.
- Emoji handling: We chose to remove emojis that appeared more than once and replaced the remaining emojis with descriptive text. This decision was based on the assumption that both the LLM and the transformer model would better interpret emojis when expressed as descriptive text rather than as emoji characters.
- Stopwords: We chose to retain the stopwords, as most of the models we plan to use are LLMs, and we believe they can identify stopwords and utilize them as contextual information.

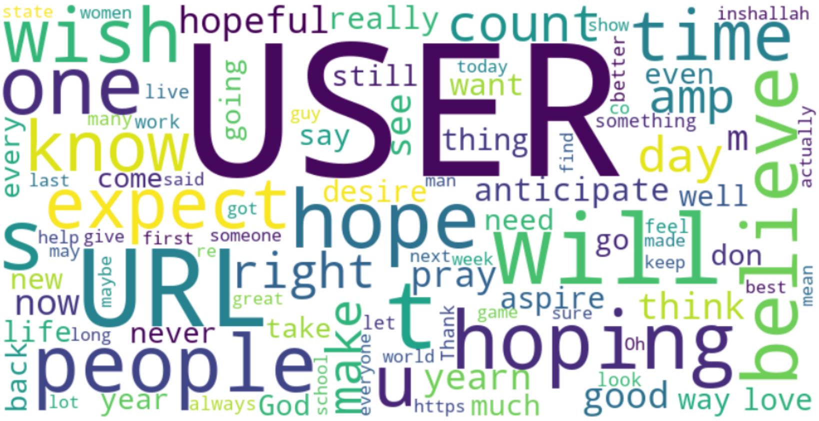


Figure 3: English Training Set

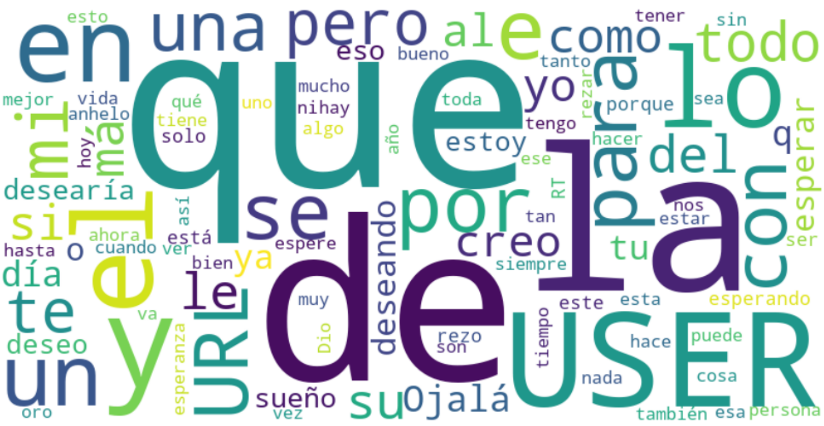


Figure 4: Spanish Training Set

3.1.2. Data augmentation

As previously mentioned, a class imbalance was observed between the Spanish and English languages. To address this, we employed data augmentation by translating the Spanish training data into English and adding it to the original English training set. Our approach was primarily guided by the number of examples in the least frequent class in the combined dataset. In the binary classification, the final distribution with augmentation was: Not Hope (5606) and Hope (4434). For the multiclass classification, the distribution was: Not Hope (4500), Generalized Hope (1385), Unrealistic Hope (1385), Realistic Hope (1385), and Sarcasm (1385).

To translate the messages, we ran the Helsinki-NLP pre-trained model with a MarianMT tokenizer from the HugginFace library in a freetier Google Colab environment with GPU support. Specifically, we used the *"Helsinki-NLP/opus-mt-es-en"* for Spanish to English model. We decided to use this HuggingFace model due to its easy implementation and fast inference.

3.2. Methods

3.2.1. XLM-RoBERTa

For the XLM-RoBERTa model, we converted the labels into numerical values and used a merged training set that combined both languages. The training parameters used were: number of train epochs: 3, learning rate: 1e-5, and max sequence length: 64. These parameters were selected through trial and error, given the limited computational resources available. We utilized Google Colab with a GPU, but due to constraints on the number of available GPU units, we were limited by the parameters allowed in this configuration. Nevertheless, the parameters were primarily based on those used in the study presented by [28].

3.2.2. Large Language Models and Prompt Engineering

The selection of LLMs for this study was guided by a combination of prior research and practical considerations. Our primary inspiration came from the work of [23], which introduced and evaluated GPT and Llama models, highlighting their effectiveness for text classification tasks. To complement these models, we included Claude, based on positive personal experience with its performance, and DeepSeek, due to its recent surge in popularity within the NLP community. This diverse selection aimed to cover models with different architectures, training data, and inference behaviors, providing a comparative perspective across leading LLM families.

Originally, our goal was to utilize the official APIs of these models to perform classifications. However, token limitations and access constraints led us to employ their respective user interfaces or chat versions instead. For each subtask and language, we defined a specific prompt, which is detailed in the next subsection. As showed in Table 1, depending on the chat capabilities of each model, we provided batches of the dataset for classification.

Model	Version	Number of texts per batch
GPT	o3-mini	200
Claude	3.5 Sonnet	200
Llama	4	60
DeepSeek	R1	100

Table 1
Overview of batch sizes used during model classification

3.2.3. Zero-Shot Prompts

For the zero-shot prompts, we adopted a unified approach, using the same prompt for binary classification in both Spanish and English. Similarly, a single prompt was designed for the multiclass classification task across both languages. This decision was made under the assumption that the model would generalize the task regardless of the input language. To further assist the model, we included the class descriptions provided on the contest page directly within the prompt for clearer guidance. The prompts used are shown below.

Binary Classification Prompt

Below, there is a list of lines of text. Your job is to decide whether the given text reflects hope or lack of hope by classifying it as either Hope or Not Hope. The definitions are:
-Hope: Hope is a crucial human emotion that influences decision-making, resilience, and social interactions.
-Not Hope: Not Hope is a text that do not express hope.
Please, give the answer in the format "number, classification". Don't forget the comma instead of a dot in your answer
Text to classify

Multiclass Classification Prompt

Below, there is a list of lines of text.Your job is to classify the text as a Generalized Hope, Realistic Hope, Unrealistic Hope, Not Hope or Sarcasm. The definitions are:
- Generalized Hope: A broad sense of optimism not tied to specific outcomes.

- Realistic Hope: Expectations grounded in achievable goals. - Unrealistic Hope: Desires for outcomes that are unlikely or impossible.
- Not Hope: Not hope is that belonged to neither category above. Texts that do not express hope.
- Sarcasm: Texts that mimic hope but are sarcastic in nature.

Please, give the answer in the format "number, classification". Don't forget the comma instead of a dot in your answer
Text to classify

3.2.4. Few Shot

For the few-shot prompts, we selected three random samples from each class in the training set, creating three example shots. We opted to use separate prompts for each language, as the examples would be specific to each language. The structure of the prompt is consistent with the zero-shot classification; the only difference is the inclusion of examples with both text and labels, which vary depending on the language. The same set of examples was used across all models.

Table 2 below presents the series of experiments conducted on the development set.

Model	Classification	Zero Shot	Few Shot	Data Augmentation
XLM-RoBERTa	Binary English	-	-	-
XLM-RoBERTa	Multiclass English	-	-	-
XLM-RoBERTa	Binary Spanish	-	-	✓
XLM-RoBERTa	Multiclass Spanish	-	-	✓
ChatGPT o3-mini	Binary English	✓	✓	-
ChatGPT o3-mini	Multiclass English	✓	✓	-
ChatGPT o3-mini	Binary Spanish	✓	✓	-
ChatGPT o3-mini	Multiclass Spanish	✓	✓	-
Claude 3.5 Sonnet	Binary English	✓	✓	-
Claude 3.5 Sonnet	Multiclass English	✓	✓	-
Claude 3.5 Sonnet	Binary Spanish	✓	✓	-
Claude 3.5 Sonnet	Multiclass Spanish	✓	✓	-
Llama 4	Binary English	✓	✓	-
Llama 4	Multiclass English	✓	✓	-
Llama 4	Binary Spanish	✓	✓	-
Llama 4	Multiclass Spanish	✓	✓	-
Deepseek R1	Multiclass English	✓	✓	-
Deepseek R1	Binary English	✓	✓	-
Deepseek R1	Multiclass Spanish	✓	✓	-
Deepseek R1	Binary Spanish	✓	✓	-

Table 2
Summary of the experiments performed on the development set.

3.3. Results

We evaluated the performance of multiple Large Lenguage Models (LLMs) using the development set. For the binary classification task, we opted for accuracy and macro-averaged F1-score as evaluation metrics. The results for each model are presented in Tables 3, and 5, corresponding to the English and Spanish datasets, respectively.

XLM-RoBERTa, with and without augmentation, achieved the highest performance across both metrics and languages, surpassing our expectations. ChatGPT o3-mini ranked second on the English dataset, with negligible differences between its zero-shot and few-shot prompting strategies—although zero-shot prompting obtained a marginally higher score. Contrastingly, in the Spanish dataset, ChatGPT o3-mini and DeepSeek R1 achieved similar results: DeepSeek R1 achieved slightly higher accuracy

Model	Setting	Accuracy	F1-Score Macro
ChatGPT o3-mini	Zero Shot	0.7992	0.7984
	Few Shot	0.7950	0.7938
Claude 3.5 Sonnet	Zero Shot	0.7555	0.7540
	Few Shot	0.7766	0.7752
Llama 4	Zero Shot	0.6993	0.6989
	Few Shot	0.7093	0.7090
Deepseek R1	Zero Shot	0.7413	0.7306
	Few Shot	0.5741	0.5665
XLM-RoBERTa	-	**0.8538**	**0.8538**
XLM-RoBERTa	Data Augmentation	0.8339	0.8338

Table 3
Comparison of models under One Shot and Few Shot settings on Accuracy and F1-Score (Macro) English Binary.

Model	Setting	Accuracy	F1-Score Weighted
ChatGPT o3-mini	Zero Shot	0.4895	0.4479
	Few Shot	0.4874	0.4463
Claude 3.5 Sonnet	Zero Shot	0.5321	0.5112
	Few Shot	0.5526	0.5434
Llama 4	Zero Shot	0.4217	0.4296
	Few Shot	0.4826	0.4922
Deepseek R1	Zero Shot	0.5231	0.5147
	Few Shot	0.4942	0.4807
XLM-RoBERTa	-	**0.7266**	**0.7287**
XLM-RoBERTa	Data Augmentation	0.6546	0.6580

Table 4
Comparison of models under One Shot and Few Shot settings on Accuracy and F1-Score (Weighted) English Multiclass.

with zero-shot prompting, while ChatGPT o3-mini outperformed in terms of F1-score under few-shot prompting.

For the multiclass classification task, we evaluated models using accuracy and the F1-score with weighted averaging, to account for label imbalance. The scores can be observed for English and Spanish datasets in Tables 4 and 6, respectively. Once again, XLM-RoBERTa led in overall performance across both datasets and metrics. As expected, all models demonstrated reduced performance on the multiclass task compared to the binary classification task.

Model	Setting	Accuracy	F1-Score Macro
ChatGPT o3-mini	Zero Shot	0.6822	0.6792
	Few Shot	0.7356	0.7351
Claude 3.5 Sonnet	Zero Shot	0.7282	0.7280
	Few Shot	0.7321	0.7307
Llama 4	Zero Shot	0.6842	0.6840
	Few Shot	0.6837	0.6834
Deepseek R1	Zero Shot	0.7324	0.7287
	Few Shot	0.7233	0.7232
XLM-RoBERTa	-	**0.8545**	**0.8544**
XLM-RoBERTa	Data Augmentation	0.7150	0.7186

Table 5
Comparison of models under One Shot and Few Shot settings on Accuracy and F1-Score (Macro) Spanish Binary.

In the English dataset, Claude 3.5 Sonnet achieved the second-highest performance under few-shot

Model	Setting	Accuracy	F1-Score Weighted
ChatGPT o3-mini	Zero Shot	0.4166	0.3195
	Few Shot	0.4281	0.2668
Claude 3.5 Sonnet	Zero Shot	0.4995	0.4030
	Few Shot	0.5066	0.4304
Llama 4	Zero Shot	0.3880	0.3395
	Few Shot	0.3841	0.3530
Deepseek R1	Zero Shot	0.5242	0.4096
	Few Shot	0.4807	0.3760
XLM-RoBERTa	-	**0.7512**	**0.7047**
XLM-RoBERTa	Data Augmentation	0.8337	0.8337

Table 6
Comparison of models under One Shot and Few Shot settings on Accuracy and F1-Score (Weighted) Spanish Multiclass.

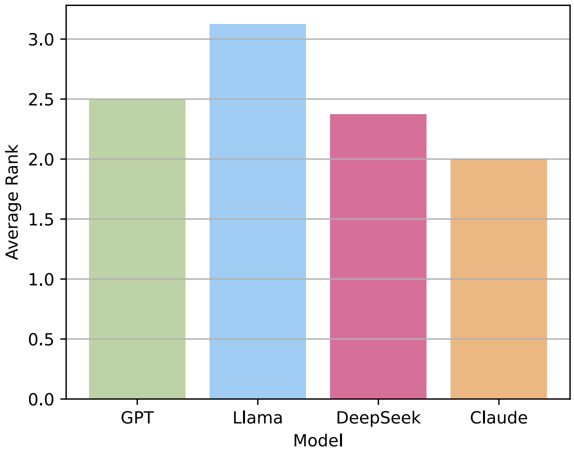


Figure 5: Average Ranks from F1-Score across LLMs. The lower ranks means a higher overall performance.

prompting. On the Spanish dataset, DeepSeek R1 (zero-shot) obtained the second-highest accuracy, while Claude 3.5 Sonnet attained the second-best F1-score.

To assess the overall performance differences among the four LLMs independently of language and prompting strategy, we applied the Friedman Test—a non-parametric statistical test commonly used for comparing the performance of multiple models across multiple datasets. Unlike ANOVA, the Friedman Test does not assume normality and is therefore well-suited for ranking-based comparisons when the same models are evaluated across several conditions.

The test evaluates the null hypothesis that all models perform equally, meaning that any observed differences in performance rankings are due to random chance. In our analysis, we obtained a p-value of 0.3691. This result indicates that we do not reject the null hypothesis, suggesting there is no statistically significant difference in performance among the evaluated LLMs. Since the null hypothesis has not been rejected, we decided to calculate a ranking based on the average F1-Scores obtained by each LLM. In Figure 5 it can be observed that Claude had the hightest rank.

In practical terms, the observed variation in scores across models and prompting strategies may be attributed to random fluctuations rather than systematic superiority of one model over another. To draw more conclusive insights, additional datasets or repeated experimental runs would be required to increase statistical power and validate any performance trends.

For the test set predictions, we submitted results from the RoBERTa model both with and without data augmentation. The version without augmentation yielded better performance. Tables 7 and 8 present a comparison between our RoBERTa without augmentation model's performance on the test set and the top five submissions in the competition. This comparison highlights how our approach measures up against the leading methods evaluated under the same conditions. Our test set scores secured a place on the leaderboard for all tasks in both English and Spanish. However, our models performed notably better on the Spanish tasks. Specifically, we ranked 15th in the English binary classification task and 9th in the English multiclass task. In contrast, we achieved 5th place in the Spanish binary task and an impressive 2nd place in the Spanish multiclass task, as shown in the table.

(a) Binary Task

Team name	Acc	Avg Mac F_1
michaelibrahim	0.8718	0.8713
rogeliorjr1	0.8706	0.8704
nayeem01	0.8706	0.8701
vandan1712	0.8681	0.8678
vishesh2002	0.8676	0.8673
dmadera	**0.8584**	**0.8583**

(b) Multiclass Task

Team name	Acc	Avg Mac F_1
supachoke	0.7815	0.7546
ebuka	0.7903	0.7484
tafredri	0.7903	0.7484
lephuquy	0.7878	0.7425
michaelibrahim	0.7827	0.7420
dmadera	**0.7563**	**0.7299**

Table 7
Comparison with top 5 results in the competition for Task 1 and Task 2 for English

(a) Binary Task

Team name	Acc	Avg Mac F_1
teddymas	0.8521	0.8521
abit7431	0.8464	0.8464
lephuquy	0.8450	0.8446
dmadera	**0.8435**	**0.8435**
supachoke	0.8376	0.8376

(b) Multiclass Task

Team name	Acc	Avg Mac F_1
lephuquy	0.7677	0.7416
dmadera	**0.7540**	**0.7221**
supachoke	0.7474	0.6984
teddymas	0.6975	0.6901
nayeem01	0.6977	0.6795

Table 8
Comparison with top 5 results in the competition for Task 1 and Task 2 for Spanish

Analyzing the binary confusion matrices on the development set using the best-performing model from the competition across the two languages, as seen in Figure 6, reveals that in the English language, the model shows a slight tendency to overpredict the "Hope" class. However, the prediction accuracy remains relatively balanced between both classes, with a slightly higher correct prediction rate for "Not Hope". In contrast, the Spanish language model exhibits a bias toward overpredicting "Hope". While the overall accuracy is the same as the English model (0.85), the Spanish model shows a stronger asymmetry in its error pattern: 153 "Hope" instances were misclassified as "Not Hope", whereas 442 "Not Hope" instances were misclassified as "Hope".

For the multiclass confusion matrices in English, the model performs best at predicting "Not Hope", with 620 correct classifications. In contrast, "Unrealistic Hope" shows the poorest performance, with only 88 correct predictions, while "Sarcasm" achieves a moderate level of accuracy with 197 correct predictions. The most common misclassification is "Not Hope" being confused with "Generalized Hope" (80 cases). For the Spanish model, the strongest performance is also in predicting "Not Hope" (1557 correct predictions), followed by a relatively strong recognition of "Generalized Hope" (780 correct predictions). The remaining three classes have lower but more balanced performance: "Unrealistic Hope" (281 correct), "Realistic Hope" (251 correct), and "Sarcasm" (202 correct). A notable pattern of confusion emerges, with "Not Hope" frequently misclassified as one of the hope-related categories.

In terms of sarcasm detection, the models show some key differences. For English, the model correctly identifies 197 instances of sarcasm but misclassifies 55 cases as other categories, with "Not Hope" being the most common misclassification (41 cases). The model also wrongly classifies 116 cases of other categories as sarcasm, with "Not Hope" (66 cases) and "Unrealistic Hope" (38 cases) being the most frequent false positives. In the Spanish model, 202 sarcasm instances are correctly identified, with 49 cases misclassified as other categories, again most frequently as "Not Hope" (33 cases). The Spanish model has 124 false positives, with "Not Hope" (59 cases) and "Unrealistic Hope" (59 cases) most

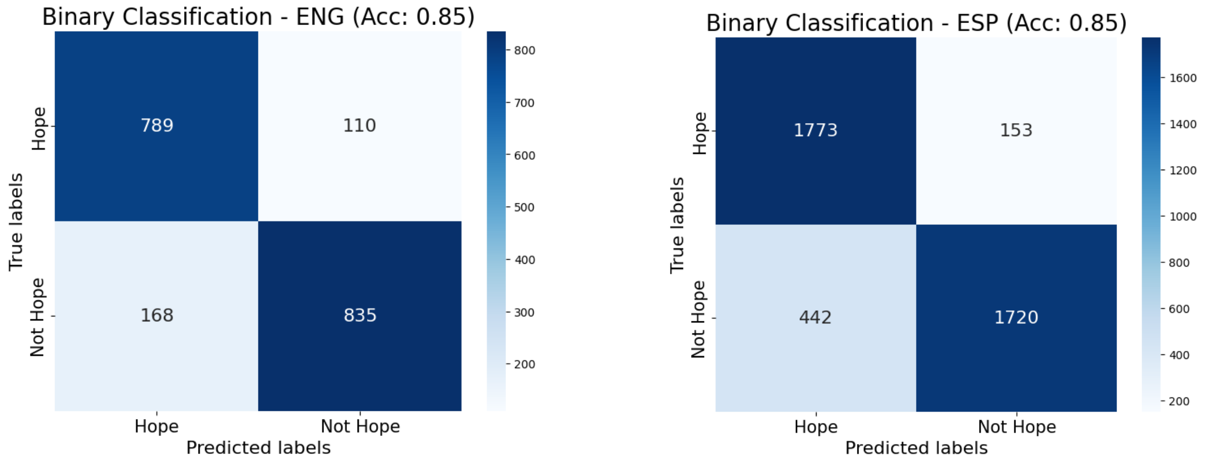


Figure 6: Confusion matrices for the binary classification task of the best model (XLM-RoBERTa without augmentation) across both languages: English (left), Spanish (right).

commonly misclassified as sarcasm.

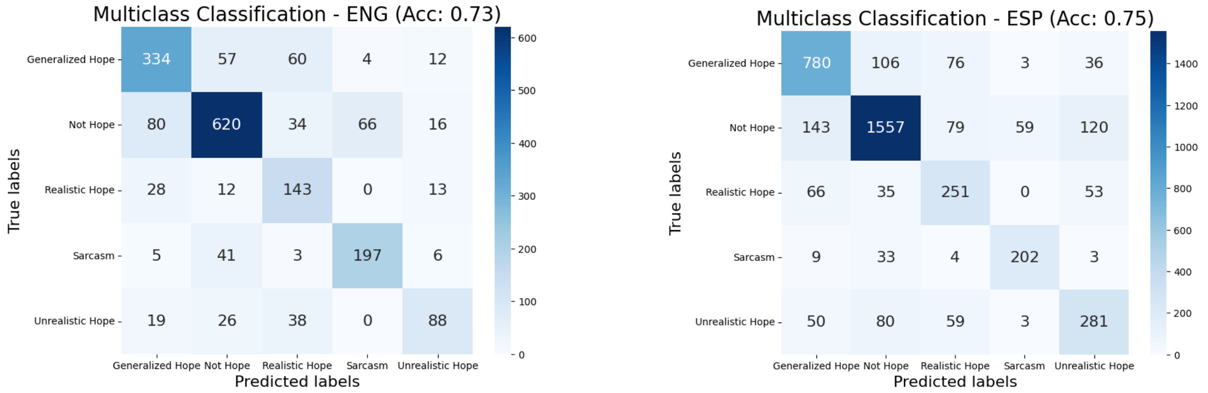


Figure 7: Confusion matrices for the multiclass classification task of the best model (XLM-RoBERTa without augmentation) across both languages: English (left), Spanish (right).

3.4. Discussion

In our experiments on hope speech and sarcasm detection across English and Spanish texts, RoBERTa consistently outperformed all tested Large Language Models (LLMs), in both binary and multiclass classification settings. This finding aligns with the results reported in related studies across other NLP tasks. For instance, SiEBERT and RoBERTa often surpass LLMs on datasets where short text length challenges LLM's accuracy [23]. Similarly, GPT-4 underperformed in sarcasm detection despite prompting strategies [27]. LLMs might show promise by offering explainable outputs and flexibility, but their performance remains generally below that of fine-tuned supervised transformers when it comes to nuanced understanding tasks such as detecting hope and sarcasm. Thus, our results further confirm that, despite their versatility, LLMs have yet to consistently surpass traditional supervised models like RoBERTa in fine-grained text classification tasks.

Furthermore, our experiments revealed interesting behaviors that are worth further analysis. In the case of ChatGPT o3-mini, its reasoning capabilities allowed us to observe parts of its internal decision-making process when classifying instances. Figure 12 illustrates several examples. In some cases, ChatGPT o3-mini relied on surface-level cues, such as the presence of the word"hoping", to make its prediction, without considering the broader context of the message. In other cases, it appeared to understand the deeper meaning, but then seemed to second-guess itself and defaulted to a simpler or

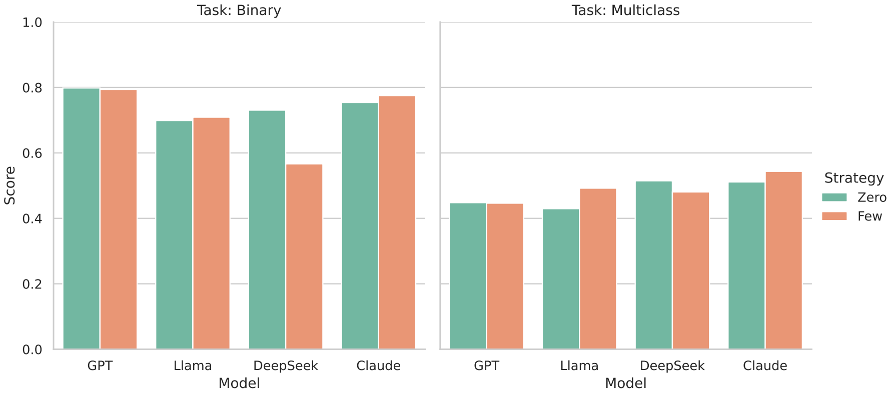


Figure 8: English: Zero-Shot vs Few-Shot Performance by Model and Task

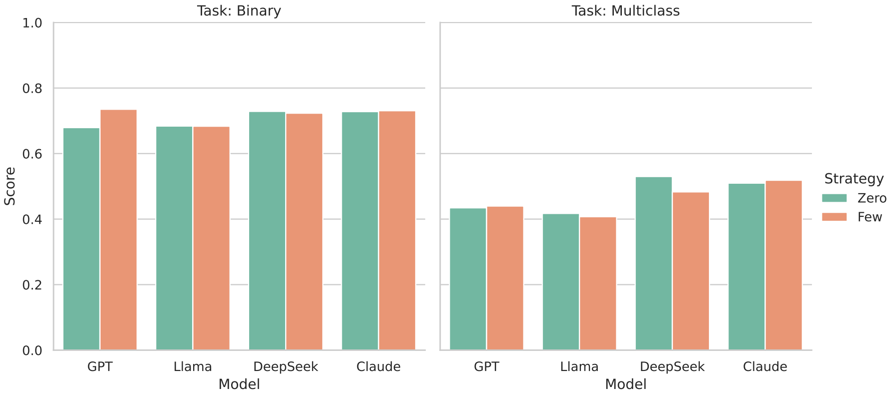


Figure 9: Spanish: Zero-Shot vs Few-Shot Performance by Model and Task

more literal interpretation, as shown in Figure 12b. We also encountered instances of hallucination where the model generated more responses than expected, even producing imaginary future texts.

With DeepSeek, the main issue we faced was censorship. In the Spanish dataset, a few instances addressed sensitive topics for the Chinese government. As a result, we had to manually removed certain proper nouns, such as "Xi," referring to the Chinese president, to avoid model refusal or skewed outputs.

Regarding the prompting strategies, we observed that the difference between zero-shot and few-shot prompting is generally minimal for the binary task. However, in the multiclass setting, almost all LLMs seem to benefit from the inclusion of examples within the instruction. Figures 8 and 9 show the scores obtained by each model under both prompting strategies for English and Spanish, respectively. In the English dataset, DeepSeek stands out as the only model that consistently performs better with zero-shot prompting, regardless of the task. A similar trend is observed in the Spanish datasets, although the difference is slightly smaller. An exception is ChatGPT, which shifts to better performance under few-shot prompting for the binary task.

On the other hand, the experiments with RoBERTa showed that data augmentation had little impact on the model's performance; in fact, it was slightly counterproductive. The results were nearly identical across both languages, as shown in Figures 10 and 11. This outcome is not surprising, given that RoBERTa is a pre-trained model with strong generalization capabilities. Notably, our test set evaluation using RoBERTa showed significantly better performance in the binary tasks for both English and Spanish when augmentation was not applied, reinforcing the idea that it was unnecessary in this case.

The confusion matrices for binary classification indicate that while both models achieve similar overall accuracy, they exhibit subtle differences in prediction biases. Specifically, the Spanish model shows a certain tendency to classify borderline cases as “Hope" compared to the English model. In contrast, the multiclass classification results reveal that both models struggle with the nuanced distinctions between

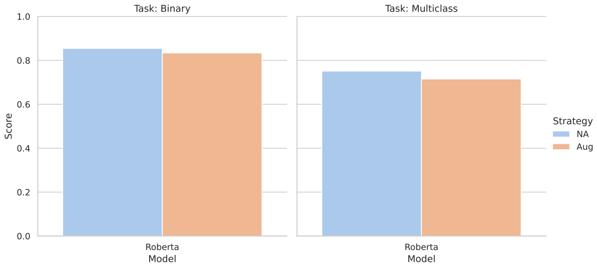


Figure 10: RoBERTa accuracy with and without using Data Augmentation during training. English data.

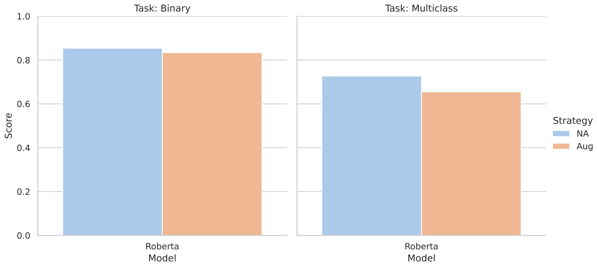


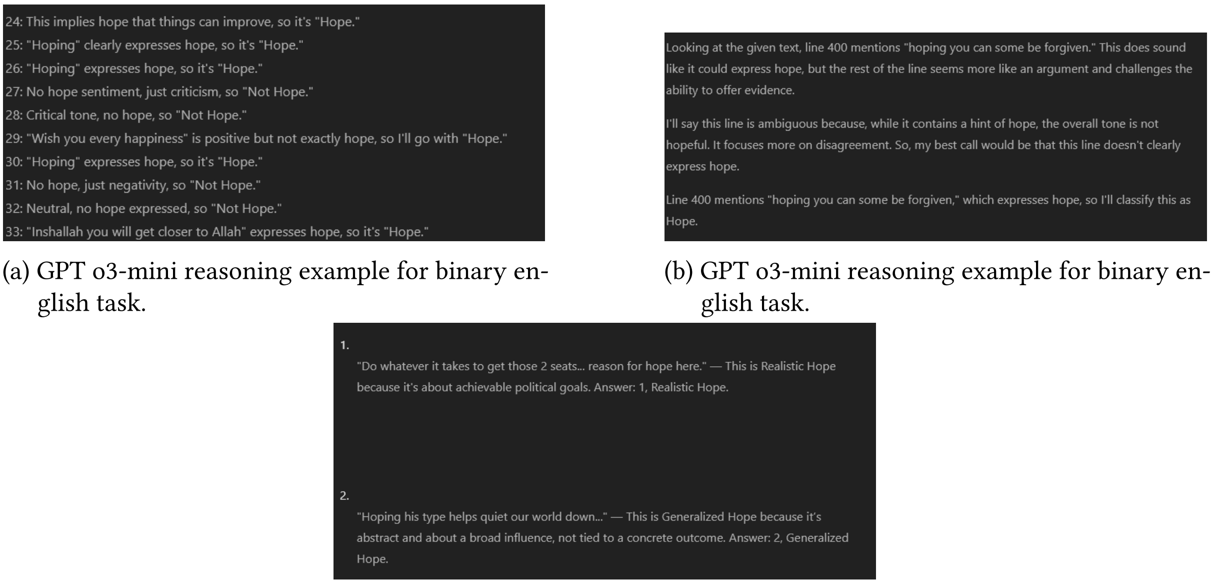
Figure 11: RoBERTa accuracy with and without using Data Augmentation during training. Spanish data.

different types of hope. However, the Spanish model demonstrates a more balanced performance across classes, whereas the English model displays a more polarized pattern, performing well in some classes but poorly in others. These findings align with the final evaluation scores, where the Spanish model emerged as the stronger overall performer.

When examining sarcasm detection across both languages, the English and Spanish multiclass confusion matrices reveal similar challenges in identifying sarcastic content. Both models struggle notably with distinguishing sarcasm from negative content, particularly “Not Hope". For the Spanish model, there is also significant confusion between sarcasm and “Unrealistic Hope". These findings suggest that while there may be some cross-language similarities in the linguistic markers of sarcasm, there are likely important differences in how sarcasm is expressed across the different categories of Hope, as its manifestation is nuanced and influenced by language-specific expressions and cultural contexts.

Plutchik’s model introduces a detailed framework of eight primary bipolar emotions, which are considered universal, innate, and essential for human survival. These basic emotions are recognized across cultures, often through distinct facial expressions, and serve as the foundation for more complex emotional experiences [29]. Recent studies on emotion classification using LLMs have shown promising results, successfully identifying emotions based on Plutchik’s wheel of emotions. These advancements are the result of combining several techniques, such as training models on diverse text datasets representing a wide range of emotional contexts, along with the use of effective prompting strategies [30]. Nonetheless, complex emotions, like hope, are shaped by conscious thought and cultural influences and often involve a mixture of feelings from different emotional categories. Because of this complexity, emotions like hope or sarcasm remain challenging to classify accurately.

For future research, we suggest evaluating the models on additional datasets to strengthen statistical validity and deepen insights into performance differences. Promoting a more balanced text distribution across multiple classes is also recommended. Although it is well-established that datasets with real-world data are inherently unbalanced, particularly when dealing with subjective topics like hope, the challenge becomes even more pronounced in emerging research areas such as sarcasm detection within the context of hope. As sarcasm remains relatively underexplored in the study of hope, the lack of sufficient labeled examples further complicates accurate classification. More annotated examples are needed to train models that can effectively distinguish between genuine expressions of hope and sarcastic ones.



24: This implies hope that things can improve, so it's "Hope."
25: "Hoping" clearly expresses hope, so it's "Hope."
26: "Hoping" expresses hope, so it's "Hope."
27: No hope sentiment, just criticism, so "Not Hope."
28: Critical tone, no hope, so "Not Hope."
29: "Wish you every happiness" is positive but not exactly hope, so I'll go with "Hope."
30: "Hoping" expresses hope, so it's "Hope."
31: No hope, just negativity, so "Not Hope."
32: Neutral, no hope expressed, so "Not Hope."
33: "Inshallah you will get closer to Allah" expresses hope, so it's "Hope."

(a) GPT o3-mini reasoning example for binary english task.

Looking at the given text, line 400 mentions "hoping you can some be forgiven." This does sound like it could express hope, but the rest of the line seems more like an argument and challenges the ability to offer evidence.

I'll say this line is ambiguous because, while it contains a hint of hope, the overall tone is not hopeful. It focuses more on disagreement. So, my best call would be that this line doesn't clearly express hope.

Line 400 mentions "hoping you can some be forgiven," which expresses hope, so I'll classify this as Hope.

(b) GPT o3-mini reasoning example for binary english task.

1.
"Do whatever it takes to get those 2 seats... reason for hope here." — This is Realistic Hope because it's about achievable political goals. Answer: 1, Realistic Hope.

2.
"Hoping his type helps quiet our world down..." — This is Generalized Hope because it's abstract and about a broad influence, not tied to a concrete outcome. Answer: 2, Generalized Hope.

(c) GPT o3-mini reasoning example for multiclass english task.

Figure 12: GPT o3 reasoning screen captures obtained during experiments.

3.5. Conclusion

Detecting complex emotions such as hope is challenging given that they arise in diverse situations with multiple emotional layers. Furthermore, distinguishing genuine hope from sarcasm presents an even more complicated task, since sarcastic comments can take the form of positive sentences to express negative feelings [5, 6]. As part of the PolyHope-M shared task at IberLEF 2025, we compare ChatGPT (o3-mini), Claude (3.5 Sonnet), Llama (4), and DeepSeek (R1), some of the most recent and popular Generative Large Language Models (LLMs) at the time of writing, against a traditional transformer-based model, RoBERTa, in detecting hope speech and sarcasm across English and Spanish texts in binary and multiclass settings.

RoBERTa consistently outperformed all tested LLMs in both binary and multiclass classification tasks in both languages, and showed no improvement using data augmentation. This demonstrates the pre-trained model's ability to detect emotions even in short-length text and its ability to generalize to other languages. Our results are also consistent with previous studies, where RoBERTa outperforms LLMs in task requiring analyzing text without much context [27, 23].

However, despite their lower classification performance, LLMs exhibited interesting and sometimes unexpected behaviors. Models such as GPT o3-mini demonstrated partial reasoning transparency, yet often defaulted to simplistic interpretations or suffered from hallucinations when handling multiple inputs. DeepSeek, while competitive in certain prompting conditions, posed unique challenges related to censorship and required careful input monitoring in specific cultural contexts.

Our evaluation of prompting strategies revealed that while the difference between zero-shot and few-shot learning was minor for binary classification, the inclusion of examples was generally beneficial in multiclass tasks. This suggests that, for more complex classification schemes, LLMs may rely heavily on the context. However, in future research, we suggest exploring other prompting techniques that make use of the generative features of the model to improve classification or use smaller input batches.

While LLMs might not yet surpass traditional transformer models such as RoBERTa in complex emotion detection, their explainability features can lead us to new insights on how these model understand emotions. Chatgpt and Deepseek showed deep reasoning when classifying a message, only to second-guess themselves and go for the more simplistic answer. For example, choosing the class “HOPE" if the word “hoping" was present, regardless of the remaining text. These situations open new research opportunities on why they can detect the emotion, but fail to follow classification instructions. Further research on these new generation of reason-focus generative large language models, or their ensembling with RoBERTa, could lead to more effective and reliable systems for hope speech and sarcasm detection in multilingual settings.

Declaration on Generative AI

Portions of this manuscript were assisted by generative AI for style and clarity. All research design, results, and interpretations are the sole work of the authors. The authors assume complete responsibility for the final content.

References

[1] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, 2022. URL: https://arxiv.org/abs/2210.14136. arXiv:2210.14136.

[2] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: M. Nissim, V. Patti, B. Plank, E. Durmus (Eds.), Proceedings of the Third Workshop on Computational Modeling of People‘s Opinions, Personality, and Emotion‘s in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5/.

[3] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, D. García-Baena, J. García-Díaz, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, P. Buitelaar (Eds.), Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022. URL: https://aclanthology.org/2022.ltedi-1.58/. doi:10.18653/v1/2022.ltedi-1.58.

[4] F. Balouchzahi, S. Butt, M. Amjad, G. Sidorov, A. Gelbukh, Urduhope: Analysis of hope and hopelessness in urdu texts, Knowledge-Based Systems 308 (2025) 112746. URL: https://www.sciencedirect.com/science/article/pii/S0950705124013807. doi:https://doi.org/10.1016/j.knosys.2024.112746.

[5] T. Dhar, S. Mishra, J. Menezes, N. Mishra, A. Alkhayyat, Exploring advanced techniques for sentiment analysis and emotion detection in social media networks, in: 2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC), 2024, pp. 1–4. doi:10.1109/ICEC59683.2024.10837473.

[6] S. Butt, F. Balouchzahi, A. I. Amjad, M. Amjad, H. G. Ceballos, S. M. Jimenez-Zafra, Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english, 2025. URL: https://arxiv.org/abs/2504.17974. arXiv:2504.17974.

[7] M. Ahmad, S. Usman, F. Humaira, A. Iqra, M. Muzzamil, A. Hmaza, G. Sidorov, I. Batyrshin, Hope speech detection using social media discourse (posi-vox-2024): A transfer learning approach, Journal of language and education 10 (2024) 31 – 43. doi:10.17323/jle.2024.22443.

[8] M. Singh, P. Motlicek, IDIAP submission@LT-EDI-ACL2022 : Hope speech detection for equality, diversity and inclusion, in: B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, P. Buitelaar (Eds.), Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 350–355. URL: https://aclanthology.org/2022.ltedi-1.54/. doi:10.18653/v1/2022.ltedi-1.54.

[9] D. García-Baena, M. Ángel García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgbt case, Language Resources and Evaluation 1 (2023) 1–28. doi:10.1007/s10579-023-09638-3.

[10] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. G. Cumbreras, A. L. Tonja, J. A. García-Díaz, S. Bozkurt, B. R. Chakravarthi, H. G. Ceballos, R. Valencia-García, G. Sidorov, L. A. U. López, A. F. Gelbukh, S. M. Jiménez-Zafra, Overview of hope at iberlef 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations, Proces. del Leng. Natural 73 (2024) 407–419. URL: https://api.semanticscholar.org/CorpusID:273550557.

[11] G. Sidorov, F. Balouchzahi, S. Butt, A. Gelbukh, Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/6/3983. doi:10.3390/app13063983.

[12] G. Sidorov, F. Balouchzahi, L. Ramos, et al., MIND-HOPE: Multilingual Identification of Nuanced Dimensions of HOPE, Preprint, Research Square, 2024. URL: https://doi.org/10.21203/rs.3.rs-5338649/v1. doi:10.21203/rs.3.rs-5338649/v1, version 1, posted 06 November 2024.

[13] S. Butt, F. Balouchzahi, J.-Z. S. M. Amjad, M., H. G. Ceballos, G. a. Sidorov, Overview of PolyHope at IberLEF 2025: Optimism, Expectation or Sarcasm?, in: Procesamiento del Lenguaje Natural, 2025.

[14] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[15] W. Mashloosh, H. Albehadili, M. Alazzawi, M. Al-Shareeda, Beyond polarity: The potential applications and impacts of sentiment analysis and emotion detection, AlKadhum Journal of Science 1 (2023) 44–51. doi:10.61710/akjs.v1i2.51.

[16] M. D. Surya Sai Eswar, N. Balaji, V. S. Sarma, Y. Chamanth Krishna, T. S, Hope speech detection in tamil and english language, in: 2022 International Conference on Inventive Computation Technologies (ICICT), 2022, pp. 51–56. doi:10.1109/ICICT54344.2022.9850866.

[17] J. M K, A. A P, KU_NLP@LT-EDI-EACL2021: A multilingual hope speech detection for equality, diversity, and inclusion using context aware embeddings, in: B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, K. Bali, P. Buitelaar (Eds.), Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 79–85. URL: https://aclanthology.org/2021.ltedi-1.10/.

[18] D. García-Baena, M. A. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish, Language Resources and Evaluation (2023) 1 – 28. URL: https://link.springer.com/content/pdf/10.1007/s10579-023-09638-3.pdf. doi:10.1007/s10579-023-09638-3.

[19] A. Dorendro, H. M. Devi, A literature review on sentiment analysis, SCRS (2024) 303 – 312. doi:10.56155/978-81-955020-9-7-29.

[20] J. Tsiligaridis, Approaches of classification models for sentiment analysis, AIRCC (2024). doi:10.5121/csit.2024.141007.

[21] B. A. Tingare, A. Jangid, Exploring the potential of transformers in natural language processing -a study on text classification, International Journal of Progressive Research in Engineering Management and Science (2024). doi:10.58257/ijprems35724.

[22] M. S. I. Malik, A. Nazarova, M. M. Jamjoom, D. I. Ignatov, Multilingual hope speech detection: A robust framework using transfer learning of fine-tuning roberta model, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101736. URL: https://www.sciencedirect.com/science/article/pii/S1319157823002902. doi:https://doi.org/10.1016/j.jksuci.2023.101736.

[23] J. O. Krugmann, J. Hartmann, Sentiment analysis in the age of generative ai, Customer Needs and Solutions 11 (2024) 3. URL: https://doi.org/10.1007/s40547-024-00143-4. doi:10.1007/s40547-024-00143-4.

[24] N. T. Thuy, D. V. Thin, An empirical study of prompt engineering with large language models for hope detection in english and spanish, in: IberLEF@SEPLN, 2024. URL: https://api.semanticscholar.org/CorpusID:273190303.

[25] H. R. LekshmiAmmal, M. Ravikiran, G. Nisha, N. Balamuralidhar, A. Madhusoodanan, A. K. Madasamy, B. R. Chakravarthi, Overlapping word removal is all you need: revisiting data imbalance in hope speech detection, Journal of Experimental & Theoretical Artificial Intelligence 36 (2024) 1837–1859. URL: https://doi.org/10.1080/0952813X.2023.2166130. doi:10.1080/0952813X.2023.2166130. arXiv:https://doi.org/10.1080/0952813X.2023.2166130.

[26] J. Zhou, An evaluation of state-of-the-art large language models for sarcasm detection, 2023. URL: https://arxiv.org/abs/2312.03706. arXiv:2312.03706.

[27] Y. Zhang, C. Zou, Z. Lian, P. Tiwari, J. Qin, Sarcasmbench: Towards evaluating large language models on sarcasm understanding, 2024. URL: https://arxiv.org/abs/2408.11319. arXiv:2408.11319.

[28] Y. Qu, Y. Yang, G. Wang, Ynu qyc at meoffendes@iberlef 2021: The xlm-roberta and lstm for identifying offensive tweets, in: IberLEF@SEPLN, 2021. URL: https://api.semanticscholar.org/CorpusID:238208186.

[29] E. Y. Chang, Behavioral emotion analysis model for large language models (invited paper), in: Proceedings of the 7th IEEE MIPR Conference, volume 14, 2024.

[30] M. Luca, G. Lopez, A. Longa, J. Kaul, How are you really doing? dig into the wheel of emotions with large language models, in: 2024 Artificial Intelligence for Business (AIxB), IEEE, 2024, pp. 72–75.