Transformers and Large Language Models for Hope Speech Detection: A Multilingual Approach

Diana P. Madera-Espíndola Tecnológico de Monterrey Estado de México, México A01025835@tec.mx
Zoe Caballero-Domínguez Tecnológico de Monterrey Estado de México, México A01747247@tec.mx
Valeria J. Ramírez-Macías Tecnológico de Monterrey Monterrey, México A01636965@tec.mx
Sabur Butt Tecnológico de Monterrey, Institute for the Future of Education Monterrey, México saburb@tec.mx
Hector G. Ceballos Tecnológico de Monterrey, Institute for the Future of Education Monterrey, México ceballos@tec.mx

Abstract

With the rise of Generative AI (GenAI) models in recent years, it is necessary to understand how they performed compared with other Deep Learning techniques, across tasks and across different languages. In this study, we benchmark ChatGPT-4 and XML-RoBERTa, a multilingual transformer-based model, as part of the Multilingual Binary and Multiclass Hope Speech Detection within the PolyHope-M 2025 shared task. Furthermore, we explored prompting techniques and data augmentation to determine which approach yields the best performance. In our experiments, XML-RoBERTa frequently outperformed ChatGPT-4. It also attained F1 scores of 0.86 for English, 0.83 for Spanish, 0.86 for German, and 0.94 for Urdu in Task 1, while achieving 0.73 for English, 0.70 for Spanish, 0.69 for German, and 0.60 for Urdu in Task 2.

1 Introduction

Hope speech detection is an emerging area in Natural Language Processing (NLP) that identifies an expectation, desire, or aspiration focused on the future, aimed at a particular or broad event or outcome, which plays a significant role in shaping human behavior, choices, and emotions (Balouchzahi et al., 2023). This task has become increasingly important in the digital age, particularly on social media platforms where content spread can contribute significantly to emotional well-being. Its relevance was especially highlighted during global crises such as the COVID-19 pandemic, in such contexts, fostering a sense of hope through language plays a crucial role in promoting resilience and mental health (Yadav et al., 2023; Surya Sai Eswar et al., 2022).

Recent research efforts have focused on of advanced machine and deep learning techniques to improve the accuracy of hope speech detection (Sidorov et al., 2024; Ahmad et al., 2024). In particular, transformer-based models have been applied to several NLP tasks and have proved a superior performance compared to other state-of-the-art models (Sidorov et al., 2023). However, the widespread adoption of large language models (LLMs) have transformed how text is represented and understood, particularly in multilingual settings (Chakravarthi, 2022; Kadiyala, 2024). This phenomenon has provoked researchers to explore the performance on sentiment analysis of new GenAI models against traditional transformer-based ones (Krugmann and Hartmann, 2024; Anas et al., 2024; Bu et al., 2024).

Despite these efforts, the experiments on the literature explore sentiment analysis broadly and there is no existing research, to the best of our knowledge, comparing GenAI and traditional models in hope classification. Therefore, in this study, we benchmarked ChatGPT-4 against XML-RoBERTa. We chose these specific models due to their popularity and performance in similar studies (Krugmann and Hartmann, 2024; Krasitskii et al., 2024; Shridhara et al., 2023). Furthermore, we explored the effectiveness of various strategies designed to optimize model performance. Specifically, we used one-shot and few-shot prompting techniques on the generative model, and data augmentation for RoBERTa.

Through a detailed evaluation of these approaches, the research provides a comprehensive analysis of how these two models compare to each other when applied to detect hope speech across diverse linguistic settings, including English, Spanish, German, and Urdu, within the framework of the PolyHope-M at RANLP 2025 shared task (Fazlourrahman et al., 2025), which emphasizes the

Our code is publicly available at https://github.com/DianaPME/PolyHope-M-RANLP-2025

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 67–76, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-008

value of harnessing existing multilingual datasets to navigate the complexities of linguistic and cultural diversity in sentiment analysis. Through this approach, it supports efforts to close communication gaps and cultivate safer, more inclusive digital communities.

2 Related Work

Social media platforms play a central role in shaping public discourse and offer a vast repository of user-generated content for linguistic analysis. These platforms provide concise and context-rich data, making them a widely used source for NLP research. Among the popular tasks explored in this domain is hate speech detection, which involves the identification and classification of language that conveys hostility, incites violence, or reinforces harmful stereotypes (Shridhara et al., 2023).

While this task aims to identify and mitigate negative online behavior, another emerging area of research is hope speech detection which serves as a source of encouragement for many people during times of illness, stress, loneliness, or depression (García-Baena et al., 2023; García-Baena et al., 2024), emphasizing the promotion of mental well-being in digital spaces (Zhu, 2022).

Relevant to this emerging task is the growing focus on diversifying the languages represented in hope speech datasets, enabling models to generalize better across linguistic and cultural contexts, support cross-linguistic transfer learning, and capture semantic nuances that vary across cultures. The HopeEDI dataset is one such effort, consisting of English, Malayalam, and Tamil YouTube comments (Chakravarthi, 2020). However, as highlighted by Gowda et al. (2022), creating effective multilingual models for hope speech detection presents substantial challenges, particularly due to language diversity and the presence of various scripts. This underscores the need for techniques such as data augmentation, including back-translation, where text is translated into another language and then back to the original to generate synthetic data. These methods are essential for expanding linguistic coverage and improving model performance in diverse language contexts (LekshmiAmmal et al., 2024).

On this line of research, the IberLEF (García-Baena et al., 2023; García-Baena et al., 2024; Butt et al., 2025) and RANLP (Sidorov et al., 2024; Balouchzahi et al., 2025) workshops on Hope Speech Detection introduce a new multilingual challenge by expanding the understanding of hope speech. It does so through the construction of a corpus that allows for both binary classification, identifying tweets as either Hope or Not Hope, and a more nuanced fine-grained categorization into three distinct types: Generalized Hope, Realistic Hope, and Unrealistic Hope. These efforts make a crucial and challenging contribution by filling a notable gap in annotated datasets dedicated to hope, since existing resources tend to omit it or misclassify it as a generic positive emotion, resulting in inaccurate predictions (Butt et al., 2025). In addition, the task provides a platform to evaluate the capabilities of advanced models in processing data across diverse linguistic contexts (Balouchzahi et al., 2022; Krasitskii et al., 2024).

In automated hope speech detection, various methods have been explored to improve performance. The introduction of transformer-based architectures has significantly impacted advancements in NLP. Models such as BERT, RoBERTa, and DistilBERT have outperformed traditional approaches, as they achieve remarkable results in a variety of applications, including hope speech detection, with the multilingual versions demonstrating the ability to effectively handling a range of languages (Dowlagar and Mamidi, 2021; Hossain et al., 2021; Sidorov et al., 2023).

On the other hand, the increasing use of Generative AI tools, particularly Large Language Models such as GPT 3 and over, has introduced promising possibilities for hope speech detection. These models can be guided using various prompting techniques, including zero-shot prompting, few-shot prompting, and chain-of-thought prompting, to generate relevant and meaningful responses (Thuy and Thin, 2024).

Since the popularization of GenAI, several researchers have been working comparing these models to the more traditional transformer-based models. Krugmann and Hartmann (2024) performed a binary and three-class sentiment classification experiment between GenAI and transformer-based models. Their experiments show that fine-tuned transfer-learning models frequently outperform general-purpose LLMs. Similarly, in an study made by Anas et al. (2024), RoBERTa attained the best performance against GenAI models in product review analysis. However, GenAI have also surpassed transformer-based models in other stud-

ies, for example, Konstantinos et al. (2024) concludes that GPT 3.5 is better at product review evaluations than BERT and RoBERTa. In another instance, ChatGPT 3.5 archived the best performance at the IberLEF 2024 hope competition for the binary task, surpassing transformer-based entries (García-Baena et al., 2024).

This experiments showcase that there is still much to learn about the use of Generative AI models for sentiment analysis, not to mention for hope detection or across languages.

3 Dataset

The dataset used in this study is sourced from the PolyHope-M dataset, which is part of the RANLP 2025 shared task https://www.codabench.org/competitions/5635/. It extends the original PolyHope dataset (Balouchzahi et al., 2023) by translating its English keywords into Spanish and German, with careful validation by native speakers to ensure linguistic and contextual accuracy. Tweets were collected using the Tweepy API and annotated by three qualified annotators per language, with final labels determined by majority vote. Additionally, also in line with the original PolyHope dataset, (Balouchzahi et al., 2025) replicated its label descriptions and definitions to develop a comparable dataset in Urdu, thereby maintaining consistency with prior work while expanding to a low-resource language. The resulting dataset is a combination of the original English, Spanish, German, and newly created Urdu data, representing the first multiclass hope speech detection dataset covering these four languages. This multilingual dataset enables comprehensive analysis and modeling of hope speech across diverse linguistic and cultural contexts, addressing a critical gap in the literature.

The data provided consists of Twitter texts in English, Spanish, German, and Urdu and is divided into three subsets: a training set, a development set, and a final test set. The development and test datasets each included three columns. One contained the tweet text, another provided the binary classification label (Hope or Not Hope), and the third represented the multiclass classification label (Generalized Hope, Realistic Hope, Unrealistic Hope, or Not Hope). In contrast, the test set included only the tweet text. It is important to note that the distribution across languages was imbalanced, with the number of Spanish and German tweets being approximately twice that of English and Urdu.

4 Methodology

4.1 Data processing

Our first step in the methodology was to clean the data to enhance the performance of the models. The text preprocessing involved standardizing the text to lowercase, trimming extra spaces, eliminating HTTP links, and removing Twitter-specific elements such as user mentions and retweet tags (rt). It also included filtering out non-alphabetical characters specific to each language, deleting emojis that appeared multiple times, and replacing the remaining emojis with their textual descriptions.

4.2 Data augmentation

As previously mentioned, a class imbalance was observed between the languages. To help mitigate this, data augmentation was applied by translating the original Spanish training data into English and the original English training data into Urdu, as only a direct translation pathway from English to Urdu was available. The resulting translated texts were then added to the respective English and Urdu training sets.

For this translation task, we used the Helsinki-NLP pre-trained machine translation model with the MarianMT tokenizer from the HuggingFace library. Specifically, we used *"Helsinki-NLP/opus-mt-en-ur"* for English to Urdu and *"Helsinki-NLP/opus-mt-es-en"* for Spanish to English. The translation was executed on a Google Colab environment with GPU support, using the free-tier account. This model was selected for its ease of implementation and efficient inference times.

4.3 XLM-Roberta

For the XLM-RoBERTa model, we converted the labels into numerical values and used a merged training set that combined all four 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 (Qu et al., 2021).

4.4 ChatGPT-4

We used the GPT-4 model that was available since 2023 but was discontinued on April 2025. Due to the limited number of available tokens, we chose to use the UI or chat versions instead. For each subtask and language, a specific prompt was defined, which will be explained in the next subsection. Furthermore, taking advantage of the model's chat capabilities, we provided the dataset in batches for classification.

4.4.1 Zero-Shot Prompts

For the zero-shot prompts, we adopted a unified approach, using the same prompt for binary classification in the four languages. Similarly, a single prompt was designed for the multiclass classification task across all 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 does 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, or Not Hope. 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.
Please, give the answer in the format "number, classification". Don't forget the comma instead of a dot in your answer
Text to classify

4.4.2 Few-Shot Prompts

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.

5 Results

We evaluated the performance of the fine-tuned XML-RoBERTa model against ChatGPT-4 in hope speech detection in four languages: English, Spanish, German, and Urdu. This evaluation was performed over two tasks: binary and multiclass. The binary task was measured using accuracy and macro-averaged F1-score as evaluation metrics, while the multiclass task used accuracy and weighted-average F1-score.

XML-RoBERTa consistently outperformed ChatGPT-4 across languages and tasks (Figure 1 and Figure 2). Table 1 shows that RoBERTa without data augmentation achieved the highest performance in both tasks in the English set. Similarly, in Spanish (Table 2, RoBERTa without augmentation again led binary classification. But the pattern breaks in multiclass classification, where RoBERTa yielded the best F1-score, while RoBERTa trained with data augmentation obtained the highest accuracy.

For the German and Urdu datasets, RoBERTa also outperformed ChatGPT-4. In German, the data augmentation version had better performance in multiclass task, while the single version in binary (Table 3). In the case of Urdu, as shown in Table

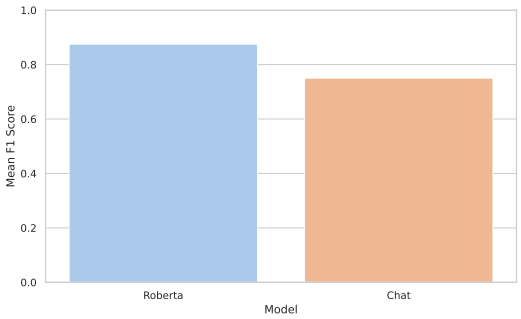

Figure 1: Average F1-score for both models, ChatGPT-4 (2023-2025) and RoBERTa, in the binary hope detection task.

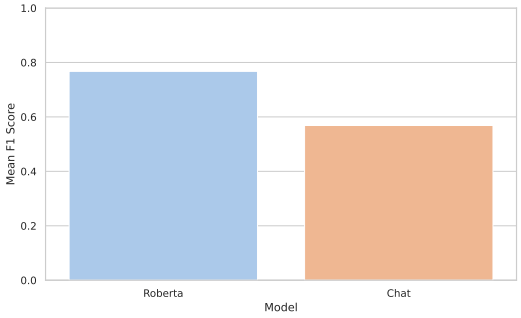

Figure 2: Average F1-score for both models, ChatGPT-4 and RoBERTa, in the multiclass hope detection task.

4) the single version slightly outperformed its augmented counterpart in both tasks. Notably, all performance differences between RoBERTa models with and without data augmentation were minimal.

Figures 3 and 4 clearly show that both models performed better on the binary classification task compared to the multiclass one. However, there is no clear trend regarding the effectiveness of zero-shot versus few-shot prompting strategies for ChatGPT-4. Although few-shot prompting yielded slightly better results in 5 out of 8 evaluations, the differences were not substantial.

Regarding the effect of data augmentation, the results suggest that RoBERTa's performance remains largely stable regardless of its inclusion. A slight advantage was noted for the non-augmented model across tasks.

For the test set predictions, we submitted RoBERTa results with and without data augmentation. Augmentation improved Spanish results, hurt Urdu performance, and had negligible impact on English and German. Tables 5, 6, 7, and 8 present a comparison between the top five places in the competition. Our scores secured a place on the leaderboard for all tasks across the four languages.

(a) Binary Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.6851	0.6860
Model 1	Few	0.7338	0.7351
Model 2	NA	**0.8433**	**0.8436**
Model 2	Aug.	0.8415	0.8418

(b) Multiclass Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.5428	0.5721
Model 1	Few	0.5453	0.5648
Model 2	NA	**0.7497**	**0.7460**
Model 2	Aug.	0.745261	0.7418

Table 1: Results on the English dataset across all models and combinations of tasks and strategies. ChatGPT-4 is denoted as Model 1; XLM-Roberta as Model 2. "Aug" refers to data augmentation applied to the training set.

(a) Binary Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.7006	0.7010
Model 1	Few	0.6949	0.6958
Model 2	NA	**0.8432**	**0.8433**
Model 2	Aug	0.8405	0.8407

(b) Multiclass Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.5095	0.4873
Model 1	Few	0.5439	0.5097
Model 2	NA	**0.7572**	0.7529
Model 2	Aug	0.7533	**0.7479**

Table 2: Results on the Spanish dataset across all models and combinations of tasks and strategies. ChatGPT-4 is denoted as Model 1; XLM-Roberta as Model 2. "Aug" refers to data augmentation applied to the training set.

Specifically, we achieved first place in both binary and multiclass tasks for English, fifth and first place for Spanish binary and multiclass tasks respectively, fourth and second place for German, and sixth and fourth place for Urdu binary and multiclass tasks respectively.

Figure 5 shows the confusion matrices obtained on the development set for RoBERTa across languages. In English, the model more accurately predicts Hope than Not Hope, but tends to misclassify Hope as Not Hope more often, reflecting a slight

Figure 3: F1 scores for zero vs few prompt strategy with ChatGPT-4 (2023-2025) across all datasets.

(a) Binary Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.7734	0.7773
Model 1	Few	0.7740	0.7785
Model 2	NA	**0.8813**	**0.8830**
Model 2	Aug	0.8778	0.8797

(b) Multiclass Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.5912	0.5682
Model 1	Few	0.5814	0.5546
Model 2	NA	0.8344	0.833888
Model 2	Aug	**0.8350**	**0.8343**

Table 3: Results on the German dataset across all models and combinations of tasks and strategies. ChatGPT-4 is denoted as Model 1; XLM-Roberta as Model 2. "Aug" refers to data augmentation applied to the training set.

(a) Binary Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.8082	0.8110
Model 1	Few	0.8323	0.8343
Model 2	NA	**0.9456**	**0.9457**
Model 2	Aug	0.9268	0.9272

(b) Multiclass Task

Model	Strategy	F1-Score	Accuracy
Model 1	Zero	0.6203	0.6227
Model 1	Few	0.6127	0.6239
Model 2	NA	**0.7547**	**0.7586**
Model 2	Aug	0.7085	0.7109

Table 4: Results on the Urdu dataset across all models and combinations of tasks and strategies. ChatGPT-4 is denoted as Model 1; XLM-Roberta as Model 2. "Aug" refers to data augmentation applied to the training set.

bias toward Hope. In the Spanish and German sets, while in Spanish RoBERTa is better at classifying Hope contrarily to German, the model makes more frequent errors misclassifying Hope than Not Hope. And Urdu shows the most balanced results.

Analyzing the multiclass confusion matrices in Figure 6, we observe distinct performance patterns across languages. In English, the model most accurately classifies Not Hope, followed by moderate success with Generalized Hope, and lower accuracy for Realistic Hope and Unrealistic Hope.

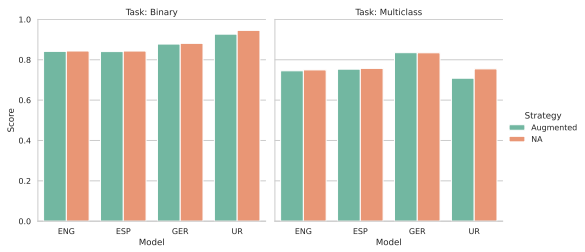

Figure 4: F1 scores for normal and data augmentation-trained RoBERTa across all datasets.

(a) Binary Task

User name	Acc	Avg Mac F_1
dmadera	**0.8634**	**0.8632**
nomanjaffar11	0.8629	0.8629
oluwatobi	0.8610	0.8608
julkarnaeen	0.8610	0.8606
teddymas	0.8557	0.8548

(b) Multiclass Task

User name	Acc	Avg Mac F_1
dmadera	**0.7801**	**0.7304**
nomanjaffar11	0.7729	0.7121
priya27	0.7680	0.7111
ahmedembedded	0.7622	0.7028
teddymas	0.7457	0.6999

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

(a) Binary Task

User name	Acc	Avg Mac F_1
nomanjaffar11	0.8499	0.8498
teddymas	0.8479	0.8478
julkarnaeen	0.8407	0.8407
priiyo9	0.8405	0.8404
dmadera	**0.8334**	**0.8326**

(b) Multiclass Task

User name	Acc	Avg Mac F_1
dmadera	**0.7660**	**0.7067**
teddymas	0.7358	0.6856
nomanjaffar11	0.7533	0.6856
abit7431	0.7377	0.6711
priiyo9	0.7433	0.6706

Table 6: Comparison with top 5 results in the competition for Task 1 and Task 2 for Spanish.

The most frequent confusions involve Generalized Hope being misclassified as Realistic Hope and as Not Hope. For Spanish, the model performs strongly on Not Hope and reasonably well on Generalized Hope, but struggles more with Realistic Hope and Unrealistic Hope. On the other hand, for German, the model excels at identifying both Not Hope and Generalized Hope, while achieving moderate accuracy on Realistic Hope and performing poorly on Unrealistic Hope. Finally, for Urdu, the model shows strong performance on Not Hope, decent accuracy on Generalized Hope and Unrealistic Hope, but severely underperforms on Realistic Hope. The most frequent misclassifications are between Generalized Hope and Unrealistic Hope

6 Discussion

The results indicate that RoBERTa consistently outperformed ChatGPT-4 across most tasks and languages, particularly in the more structured binary classification setting. These findings are consistent with prior research in related NLP tasks, where fine-tuned supervised transformers such as RoBERTa

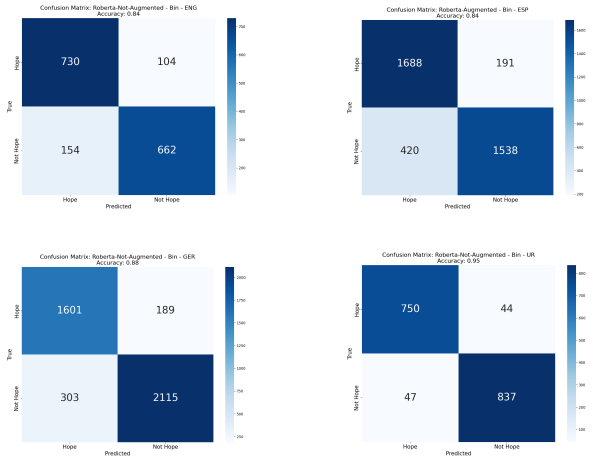

Figure 5: Confusion matrices for the binary classification task of the best model (XLM-RoBERTa) across four languages: English (top left), Spanish (top right), German (bottom left), and Urdu (bottom right).

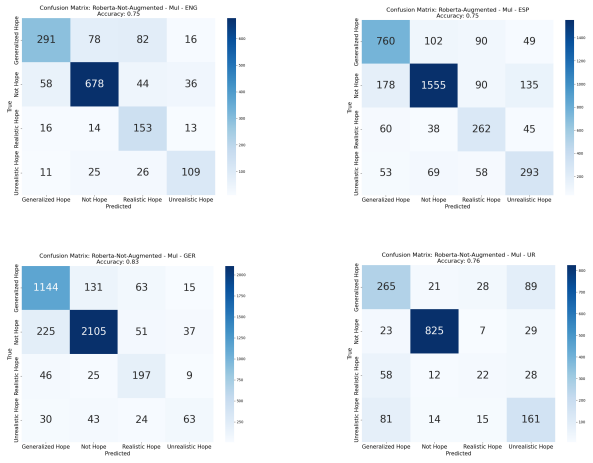

Figure 6: Confusion matrices for the multiclass classification task of the best model (XLM-RoBERTa) across four languages: English (top left), Spanish (top right), German (bottom left), and Urdu (bottom right).

(a) Binary Task

User name	Acc	Avg Mac F_1
teddymas	0.8746	0.8726
nomanjaffar11	0.8742	0.8715
abit7431	0.8668	0.8638
dmadera	**0.8647**	**0.8633**
unstoppable	0.8576	0.8568

(b) Multiclass Task

User name	Acc	Avg Mac F_1
nomanjaffar11	0.8345	0.7013
dmadera	**0.8229**	**0.6968**
teddymas	0.8135	0.6944
abit7431	0.8172	0.6778
julkarnaeen	0.8004	0.6741

Table 7: Comparison with top 5 results in the competition for Task 1 and Task 2 for German.

(a) Binary Task

User name	Acc	Avg Mac F_1
abit7431	0.9499	0.9498
nomanjaffar11	0.9499	0.9498
teddymas	0.9499	0.9498
oluwatobi	0.9480	0.9480
ahmedembedded	0.9461	0.9461
dmadera	**0.9451**	**0.9451**

(b) Multiclass Task

User name	Acc	Avg Mac F_1
nomanjaffar11	0.7836	0.6526
abit7431	0.7736	0.6482
teddymas	0.7655	0.6314
dmadera	**0.7769**	**0.6079**
priiyo9	0.7636	0.6015

Table 8: Comparison with top 5 results in the competition for Task 1 and Task 2 for Urdu.

often outperform Large Language Models (LLMs) on tasks requiring nuanced understanding of short texts. For instance, Krugmann et al. (2024) and Zhang (2024) highlight that models like SiEBERT and RoBERTa excel on short-form content where LLMs tend to struggle.

Furthermore, our results are also similar to those in the baseline experiments made by Sidorov et al. (2024) and Balouchzahi et al. (2025). Table 9 shows the comparison between the baseline results obtained with RoBERTa variants presented by the authors and our Test set results. Sidorov et al. used the model *"FacebookAI/xlm-roberta-base"* for English, Spanish, and German. Balouchzahi et al. used *"urduhack/roberta-urdu-small"* for the Urdu dataset. All models were obtained through HuggingFace. Meanwhile, we used the same model for all datasets (*"xlm-roberta-large"*) and fine-tuned it using the training set plus data augmentation. As the table shows, we outperformed the baseline results, confirming that XML-RoBERTa is a powerful transformer for hope detection and that the

(a) Binary Task

Authors	ENG	ES	GER	UR
1	0.8623	0.8369	0.8704	-
2	-	-	-	0.6961
3	0.8632	0.8326	0.8633	0.9451

(b) Multiclass Task

Authors	ENG	ES	GER	UR
1	0.6907	0.6801	0.6878	-
2	-	-	-	0.4801
3	0.7304	0.7067	0.6968	0.6079

Table 9: Comparison of Avg Macro F_1 scores between baseline results from Sidorov et al. (1), and Balouchzahi et al. (2), and our proposed method (Madera et al. (3)) evaluated on the test set.

addition of data augmentation can lead to better performance.

Interestingly, our experiments showed a strong performance in under-resourced languages such as Urdu and German, an unexpected outcome given that English and Spanish are more prominently represented in large-scale datasets and benchmarks (Balouchzahi et al., 2025). The binary confusion matrices indicate that linguistic features or language-specific training data characteristics influence how the model allocates predictions between the two classes, with German and Spanish showing the strongest biases toward 'Not Hope' compared to English and Urdu. The fact that both models maintained reasonable effectiveness across these languages suggests that multilingual models like XLM-RoBERTa can successfully transfer knowledge to underrepresented languages. However, further investigation is needed to confirm this trend and to ensure equitable performance across diverse linguistic contexts.

In contrast, for the multiclass classification task, all models perform best at predicting 'Not Hope', with the exception of German, where the model excels. Across Spanish, German, and Urdu, 'Realistic Hope' consistently emerges as the most challenging class to predict. This multiclass analysis highlights the model's difficulties in distinguishing nuanced hope categories across languages, with each language exhibiting distinct patterns of confusion between specific class pairs.

It is important to note that we trained RoBERTa with scarce computational resources and a short time-period. Therefore, while we obtained superior results, these can be improve with prolonged training or further hyperparameter optimization.

Finally, regarding ChatGPT-4, we recommend exploring additional prompting techniques and test in smaller batch settings. Generative AI has great potential for sentiment analysis and its continuous growth in use (Kim, 2024), including cases of emotional companionship, justifies the need for continued research on how they can detect complex emotions such as hope.

7 Conclusion

Hope speech detection is a growing field in NLP that seeks to identify expressions of expectation, aspiration, or encouragement. These emotions have a key role on human behavior and emotional well-being (Balouchzahi et al., 2023). In the present study, we evaluate and compare the effectiveness of two approaches: transformer-based model XML-RoBERTa with and without data augmentation, and the generative large language model ChatGPT-4 using zero-shot and few-shot prompting. We use the multilingual dataset provided by the PolyHope-M shared task at RANLP 2025, and assess both binary and multiclass classification tasks across English, Spanish, German, and Urdu.

The results demonstrate that RoBERTa consistently outperformed ChatGPT-4 across all tasks and languages, with notable higher performance in the binary classification setting. These findings support prior evidence that supervised model remain highly effective for short-text emotion detection, while LLMs may struggle due to their context dependence. For future work, we suggest exploring other prompts that leverage the LLMs generative abilities for better classification, as well as further hyperparameter optimization for RoBERTa models.

References

M.A. Ahmad, Sardar Usman, Farid Humaira, Ameer Iqra, Muhammad Muzzamil, Ameer Hmaza, Grigori Sidorov, and Ildar Batyrshin. 2024. Hope speech detection using social media discourse (posi-vox-2024): A transfer learning approach. *Journal of language and education*, 10:31 – 43.

Mohammad Anas, Anam Saiyeda, Shahab Saquib Sohail, Erik Cambria, and Amir Hussain. 2024. Can generative ai models extract deeper sentiments as compared to traditional deep learning algorithms? *IEEE Intelligent Systems*, 39(2):5–10.

Fazlourrahman Balouchzahi, Sabur Butt, Maaz Amjad, Grigori Sidorov, and Alexander Gelbukh. 2025. Urduhope: Analysis of hope and hopelessness in urdu texts. *Knowledge-Based Systems*, 308:112746.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2022. Polyhope: Two-level hope speech detection from tweets.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Kun Bu, Yuanchao Liu, and Xiaolong Ju. 2024. Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, 283:111148.

Sabur Butt, Fazlourrahman Balouchzahi, Ahmad Imam Amjad, Maaz Amjad, Hector G. Ceballos, and Salud Maria Jimenez-Zafra. 2025. Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Suman Dowlagar and Radhika Mamidi. 2021. EDIOne@LT-EDI-EACL2021: Pre-trained transformers with convolutional neural networks for hope speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 86–91, Kyiv. Association for Computational Linguistics.

Balouchzahi Fazlourrahman, Sabur Butt, Maaz Amjad, Luis Jose Gonzalez-Gomez, Abdul Gafar Manuel Meque, Helena Gomez-Adorno, Bharathi Raja Chakravarthi, Grigori Sidorov, Thomas Mandl, Ruba Priyadharshini, Hector Ceballos, and Saranya Rajiakodi. 2025. Overview of PolyHope-M at RANLP: Bridging Hope Speech Detection Across Multiple Languages. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*.

Daniel García-Baena, Fazlourrahman Balouchzahi, Sabur Butt, Miguel Ángel García Cumbreras, Atnafu Lambebo Tonja, José Antonio García-Díaz, Selen Bozkurt, Bharathi Raja Chakravarthi, Hector G. Ceballos, Rafael Valencia-García, Grigori Sidorov, Luis Alfonso Ureña López, Alexander F. Gelbukh, and Salud María Jiménez-Zafra. 2024. 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*, 73:407–419.

Daniel García-Baena, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Hope speech detection in spanish: The lgbt case. *Language Resources and Evaluation*, 1:1–28.

Anusha Gowda, Fazlourrahman Balouchzahi, H. L. Shashirekha, and Grigori Sidorov. 2022. Mucic@lt-edi-acl2022: Hope speech detection using data resampling and 1d conv-lstm. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 161 – 166. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@lt-edi-eacl2021: Multilingual code-mixed hope speech detection using cross-lingual representation learner. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 168 – 174. Association for Computational Linguistics.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual emotion detection through large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 464–469, Bangkok, Thailand. Association for Computational Linguistics.

Minseong Kim. 2024. Unveiling the e-servicescape of chatgpt: Exploring user psychology and engagement in ai-powered chatbot experiences. *Behavioral Sciences*, 14:558.

Mikhail Krasitskii, Olga Kolesnikova, Liliana Hernandez, Grigori Sidorov, and Alexander Gelbukh. 2024. Hope2024@iberlef: A cross-linguistic exploration of hope speech detection in social media. In *40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, volume 459, pages 407–419.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Hariharan RamakrishnaIyer LekshmiAmmal, Manikandan Ravikiran, Gayathri Nisha, Navyasree Balamuralidhar, Adithya Madhusoodanan, Anand Kumar Madasamy, and Bharathi Raja Chakravarthi. 2024. Overlapping word removal is all you need: revisiting data imbalance in hope speech detection. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(8):1837–1859.

Yuanchi Qu, Yanhua Yang, and Gang Wang. 2021. Ynu qyc at meoffendes@iberlef 2021: The xlm-roberta and lstm for identifying offensive tweets. In *IberLEF@SEPLN*.

Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. 2024. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Manohar Gowdru Shridhara, Viktor Pristaš, Albert Kotvytskiy, L'ubomír Antoni, and Gabriel Semanišin. 2023. A short review on hate speech detection: challenges towards datasets and techniques. In *2023 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 204–209.

Grigori Sidorov, Fazlourrahman Balouchzahi, Sabur Butt, and Alexander Gelbukh. 2023. Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets. *Applied Sciences*, 13(6).

Grigori Sidorov, Fazlourrahman Balouchzahi, Luís António de Oliveira Ramos, Helena Gómez-Adorno, and Alexander Gelbukh. 2024. Mind-hope: Multilingual identification of nuanced dimensions of hope. *Research Square*.

Medicharla Dinesh Surya Sai Eswar, Nandhini Balaji, Vedula Sudhanva Sarma, Yarlagadda Chamanth Krishna, and Thara S. 2022. Hope speech detection in tamil and english language. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 51–56.

Nguyen Thi Thuy and Dang Van Thin. 2024. An empirical study of prompt engineering with large language models for hope detection in english and spanish. In *IberLEF@SEPLN*.

Neemesh Yadav, Mohammad Aflah Khan, Diksha Sethi, and Raghav Sahni. 2023. Beyond negativity: Reanalysis and follow-up experiments on hope speech detection.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2024. Sarcasmbench: Towards evaluating large language models on sarcasm understanding.

Yue Zhu. 2022. LPS@LT-EDI-ACL2022:an ensemble approach about hope speech detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 183–189, Dublin, Ireland. Association for Computational Linguistics.