SupaChoke at IberLEF2025 PolyHope: Custom BGE Models For Multilingual Hope Speech Detection

Nguyen Phu Thanh[1,2,*], Cu Nguyen Huy Thai Tuan[1,2,*] and Nguyen Trong Chinh[1,2,†]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

Hope is a fundamental human emotion that significantly influences behavior, mood, and decision-making. Its nuanced nature—especially when combined with figurative language like sarcasm—presents challenges for Natural Language Processing (NLP) systems. In this paper, we present our submissions to the PolyHope shared task at IberLEF 2025, which focuses on detecting hope speech in tweets in English and Spanish. The task consists of Binary Classification and Multiclass Classification for each language. To address these challenges, we propose a robust approach based on transformer-based models, utilizing BGE (BAAI General Embedding) architecture. Our results demonstrate the effectiveness of our approach: we placed 11th in English Binary Classification, 1st in English Multiclass Classification, 5th in Spanish Binary Classification, and 3rd in Spanish Multiclass Classification, based on averaged F1 scores. These results highlight the power of advanced multilingual transformer architectures in addressing nuanced affective classification tasks in social media content. Our source code is published on https://github.com/NPTIsMyName/supachoke-PolyHope2025

Keywords

Hope Speech Detection, Multilingual, BERT-based model, Sentiment Classification, IberLEF

1. Introduction

The PolyHope shared task at IberLEF 2025 [1] focuses on analyzing the expression of hope in social media texts, e.g. tweets [2, 3], inspired by the goals and methodologies of previous shared tasks on hope speech detection [4, 5, 6]. While hope is a fundamental human emotion influencing behavior, mood, and decision-making, its nuanced nature [7, 8], especially when masked by sarcasm, poses challenges for Natural Language Processing (NLP) systems. This edition introduces novel dimensions, including differentiating genuine hope from sarcasm and expanding the study to English and Spanish texts, focusing on hope as an expectation. In this shared task, two sub-tasks were proposed for participants. The first challenge, called Binary Hope Speech Detection, aims to classify whether a given text in English or Spanish conveys hope. For example, given a new English tweet, "I believe things will get better soon, we just have to stay strong.", based on the presence of hopeful sentiment, either directly expressed or subtly implied, the output for this task should be "Hope". On the other hand, the second task, called Multiclass Hope Speech Detection, focuses on identifying the type of hope being expressed. If a tweet is labeled as "Hope", it is further classified into one of the following categories: “Generalized Hope”, “Realistic Hope”, “Unrealistic Hope”, or “Sarcasm” [9, 10]. Otherwise, it is classified as “Not Hope”, depending on the nature and tone of the expression.

This study presents a comprehensive approach to hope classification using various BERT-based NLP models, with a particular focus on the performance of two BGE models—BGE-m3 [11] and BGE-large-en-v1.5 [12]—across two classification subtasks. Specifically, we apply BGE-m3 for the Spanish task and BGE-large-en-v1.5 for the English task, leveraging their respective language optimizations. We also explore different fine-tuning strategies to assess their impact on performance.While the study covers

IberLEF 2025 September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
23521452@gm.uit.edu.vn (N. P. Thanh); 23521706@gm.uit.edu.vn (C. N. H. T. Tuan); chinhnt@uit.edu.vn (N. T. Chinh)
https://github.com/NPTIsMyName (N. P. Thanh); https://github.com/thaituanUIT (C. N. H. T. Tuan)

both languages and subtasks, the later sections of the paper place particular emphasis on the English multiclass classification, which was our primary focus and gained the best results.

The paper is organized as follows: Section 2 describes the proposed methodology. Section 3 outlines the experimental workflow. Section 4 presents and discusses the results. In Section 5, we provide an error analysis of the model's predictions, and finally, Section 6 concludes with a summary of our findings and future research directions.

2. Related Works

Hope, fundamentally, embodies an individual's aspiration for a particular outcome coupled with a belief in its attainability, reflecting both a desire and a cognitive commitment to pursue objectives despite inherent uncertainties [13]. This construct has garnered significant scholarly attention across disciplines, including psychology, sociology, and, increasingly, computational linguistics [14]. Understanding how hope is expressed in language, especially in social media and digital communication, allows researchers to build models that have the capability of detecting and interpreting hopeful speech. In recent years, a growing body of research has explored this phenomenon across languages, platforms, and cultural contexts, employing traditional machine learning techniques and deep learning approaches. Several studies have investigated the recognition of hope in language.

Sidorov et al. [15] conducted a comprehensive analysis of transformer models on the task of detecting regret and hope speech, utilizing two datasets: ReDDIT for regret and PolyHope for hope. Their results showed that RoBERTa achieved the highest average F1-score (0.83) for regret detection, while uncased BERT led in hope detection with an F1-score of 0.72. These findings underscore the importance of model architecture, pretraining strategy, and contextual embeddings in handling nuanced emotional content.

In HOPEIberLEF 2024, Krasitskii et al. [16] presented a notable contribution to the hope speech detection expanded across linguistic and cultural boundaries. Their work explored hope speech in English and Spanish social media using transformer-based models such as BERT, addressing binary and multiclass classification challenges. This study emphasized the importance of multilingual approaches and nuanced annotation schemes in identifying constructive and empathetic online discourse.

In their participation at the HOPEIberLEF 2024 shared task, Ronghao Pan et al. [17] addressed the challenge of detecting hope speech in social media, focusing on two perspectives: hope related to equality, diversity, and inclusion (EDI), and hope as an expression of expectation. Their approach involved fine-tuning pre-trained Transformer-based models, integrating outputs from sentiment and emotion identification models to enhance the detection of hopeful language. This methodology enabled the models to better capture the emotional context of hope speech. The UMUTeam achieved competitive results, ranking eighth in Task 1 with an M-F1 score of 0.60, and performed among the top teams in other subtasks, including Task 2.a for both Spanish and English datasets. Notably, their approach demonstrated consistent performance across various tasks without relying on data augmentation or complex model ensembles, highlighting the significance of incorporating emotional and sentiment features in hope speech detection.

3. Methodology

3.1. Data Overview & Preprocessing

3.1.1. Data Overview

The original dataset [18] includes tweet comments in both English and Spanish. Table 1 below provides statistics for the training data, showing the distribution between two binary classes and multiple classes in the training and evaluating dataset, respectively.

During the data overview, we observed that both language training datasets contained numerous unnecessary tokens such as stopwords, misspellings, slang, emojis, email addresses, URLs, and other

Category	English	Spanish
Binary		
Hope	2,426	5,316
Not Hope	2,807	5,927
Multiclass		
Realistic Hope	540	1,113
Unrealistic Hope	472	1,300
Generalized Hope	1,284	2,754
Not Hope	2,245	5,383
Sarcasm	692	693

(a) Training Dataset

Category	English	Spanish
Binary		
Hope	1,003	1,926
Not Hope	899	2,162
Multiclass		
Realistic Hope	196	405
Unrealistic Hope	171	473
Generalized Hope	467	1,001
Not Hope	816	1,958
Sarcasm	252	251

(b) Evaluating Dataset

Table 1
Statistics of the training and evaluating datasets for both English and Spanish.

irrelevant strings (e.g., “ahagdha,” “bfgyeyd”), primarily caused by user typographical errors. Additionally, placeholders (e.g., "#USER#") were frequently encountered. These noisy tokens significantly complicate the data preprocessing.

3.1.2. Data Preprocessing

Initially, in the Development phase of the competition, we distinguished the original training dataset provided by the organizers into two subsets, namely “partly clean” set and “completely clean” set. In the “partly clean” set, we cleaned the data by converting all text to lowercase, removing emojis, URLs, punctuation, and trimming unnecessary whitespace. On the other hand, in the “completely clean” set, we applied the same basic cleaning techniques and additionally performed stop-word removal and lemmatization.

In the Test phase, since we didn’t utilize any external datasets, we concatenated the “train” and “dev” datasets into a single training set to increase the amount of data available for training our models, which we expected to potentially improve inference performance. After merging, the new dataset was again distinguished into two sets, “partly clean” and “completely clean”, following the same pre-processing criteria as mentioned in the Development phase.

3.2. Model Construction & Findings

3.2.1. Traditional Deep Learning Approach

To address the binary classification task in the Development phase, we initially developed a deep learning model based on a Bidirectional Long Short-Term Memory (Bi-LSTM) [19] and a simple RNN architecture, implemented using the Sequential API from the TensorFlow framework. The model structure is depicted in Figure 1. The training and evaluation datasets in this phase consisted of text samples paired with corresponding labels, which were encoded using Scikit-learn’s LabelEncoder to represent either hope or non-hope speech. In this approach, we utilized feature extraction with word embeddings derived from FastText [20], specifically leveraging the pre-trained word2vec-google-news-300 vectors. The model was compiled using the Adam optimizer and binary cross-entropy as the loss function. Training was conducted over 10 epochs. Despite the implementation of this deep learning architecture, the model’s predictive performance remained relatively low on both the “partly clean” and “completely clean” datasets. The best F1-score achieved by our Bi-LSTM model was 0.46 on the “partly clean” dataset, and only 0.57 on the “completely clean” dataset, and RNN’s scores were only 0.38 and 0.41, indicating limited generalization capacity in comparison to later transformer-based models as discussed in the following sections.

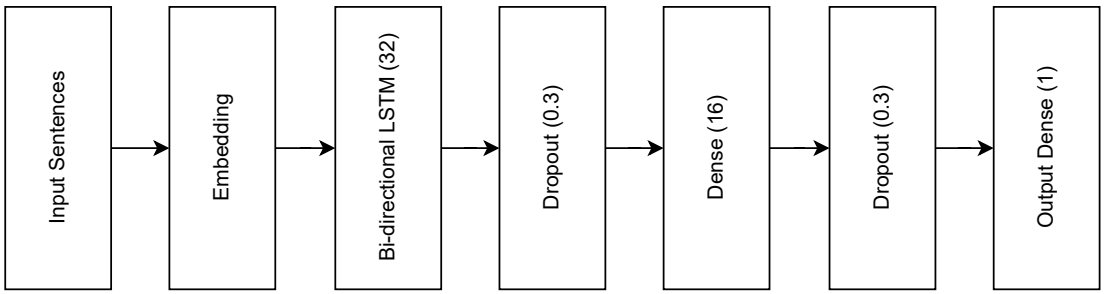


Figure 1: The architecture of Bi-LSTM

3.2.2. BERT-based Models

BERT-based models have succeeded in many NLP tasks [21], which is the reason we took the BERT-based model as our main core and applied some hyperparameter modifications to them. In the next approach, Based on an extensive review of recent research on BERT architecture and its variants, we identified and selected three State-Of-The-Art transformer-based models for our classification task, namely DeBERTa-v3-large [22], RoBERTa-large [23], BGE-m3 [24] and BGE-large-en-v1.5 [25], a BERT-based model developed by the Beijing Academy of Artificial Intelligence—that are pre-trained on multilingual and sentiment-focused datasets, these models were chosen due to their consistently strong performance as reported in prior benchmark studies and peer-reviewed publications. To evaluate the performance of different pre-trained language models on our task, we experimented on both "partly clean" and "completely cleaned" versions of the dataset, using the same training configuration. By applying identical hyperparameters and training strategies across these models, we ensured a fair comparison to assess their relative effectiveness on binary and multiclass classification tasks of both languages.

3.2.3. Experimenting

All models we experimented with were trained using the same set of hyperparameters to ensure a fair comparison across different transformer architectures. The entire pipeline was implemented in Python using the PyTorch deep learning framework, with HuggingFace's Transformers and Datasets libraries for model integration and data preprocessing. Prior to training, the input data was tokenized using the corresponding pretrained tokenizer from the HuggingFace Hub for each model architecture [26, 27, 11, 12]. The tokenized data was then encoded into input-ids, attention-mask, and token-type-ids (when required), forming the standard input structure expected by transformer-based models. Feature vectors were generated through BERT-style contextual embeddings derived from the encoder outputs. The models were trained using a custom pipeline consisting of data loading, tokenization, model instantiation, training loop, and evaluation, orchestrated using the Trainer API from HuggingFace. Training was conducted on two NVIDIA T4 GPUs provided by Kaggle Notebooks, with each model completing training within approximately three hours. After training, the models were saved as pretrained checkpoints for reproducibility and later inference.

The following training configuration was used throughout the experiments:

- **Learning rate:** 1e-5
- **Batch size:** 4 per device (for both training and evaluation)
- **Gradient accumulation steps:** 4
- **Number of epochs:** 5
- **Warm-up steps:** 200
- **Optimizer:** AdamW (PyTorch implementation)
- **Weight decay:** 0.01

- **Precision:** Mixed precision training enabled (`fp16=True`)
- **Gradient checkpointing:** Enabled to reduce memory usage
- **Random seed:** 221 (for reproducibility)
- **Logging:** The metrics were recorded at the end of each epoch

Epoch	Partly Clean		Completely Clean	
	Val Loss	F1	Val Loss	F1
1	0.8766	0.5930	0.9572	0.4940
2	**0.7478**	0.6792	0.7934	0.6681
3	0.7737	0.7013	**0.7535**	0.6903
4	0.9113	0.7090	0.8484	**0.6916**
5	1.1546	**0.7107**	1.0678	0.6868

(a) DeBERTa-v3-large

Epoch	Partly Clean		Completely Clean	
	Val Loss	F1	Val Loss	F1
1	0.9752	0.5042	0.8564	0.6171
2	0.7140	0.6926	**0.7250**	0.6913
3	**0.7114**	0.7133	0.7363	0.6979
4	0.7788	**0.7178**	0.8136	0.6847
5	0.9990	0.7175	0.9836	**0.7019**

(b) RoBERTa-large

Epoch	Partly Clean		Completely Clean	
	Val Loss	F1	Val Loss	F1
1	0.9011	0.5287	0.7867	0.6332
2	0.7414	0.6932	**0.6977**	**0.7049**
3	**0.6730**	0.7090	0.7658	0.6887
4	0.7365	**0.7214**	0.9430	0.7013
5	0.8453	0.7168	—	—

(c) BGE-large-en-v1.5

Table 2
Performance Comparison of BERT-based Models on Partly Clean vs. Completely Clean Datasets (Multiclass English)

Note: The fifth epoch of the BGE-large-en-v1.5 (Completely Clean) model is omitted due to early stopping, which was triggered after two consecutive epochs without improvement in validation performance.

After evaluating the performance of three BERT-based models: DeBERTa-v3-base, BGE-large-en-v1.5 and RoBERTa-large on both "partly clean" and "completely clean" English datasets, we observed that some techniques, such as lemmatization and stop-word removal, did not improve the model's F1-score; in some cases, they even led to performance degradation. Thus, we excluded stop-word removal and lemmatization from our final pre-processing pipeline. Among the evaluated models, BGE-large-en-v1.5 consistently outperformed the others on the Multiclass English task, achieving a lower validation loss and a notably higher F1-score. Based on this observation, we selected BGE-large-en-v1.5 as the primary model for training across the remaining tasks, including Binary English, Binary Spanish, and Multiclass Spanish.

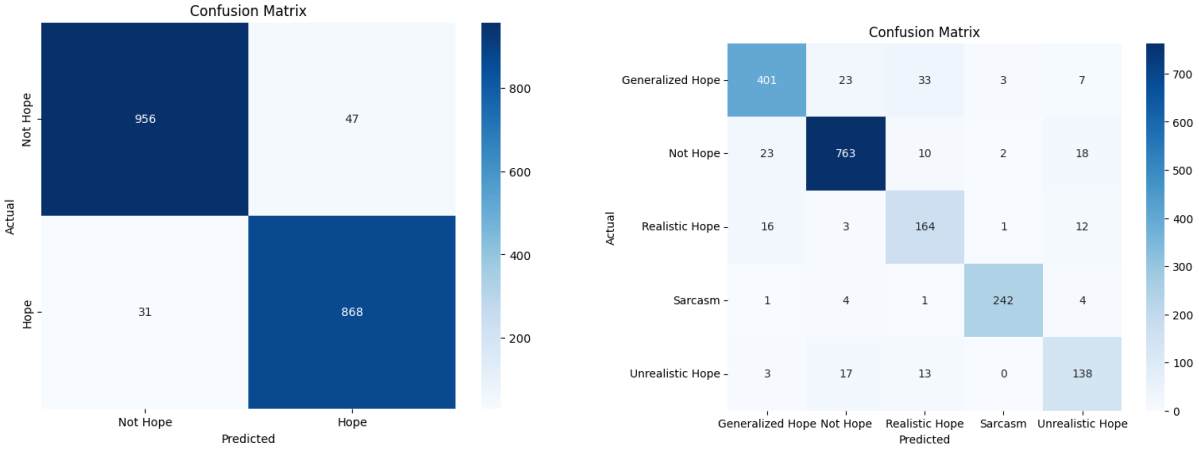
During experiments with the English Multiclass task, we noticed that the validation loss plateaued at epoch 4 despite the training loss continuing to decrease. Based on this observation, we hypothesized that the learning rate of 1e-5 might be too large for further fine-tuning beyond this point. Rather than relying on automated learning rate schedulers, we adopted a manual adjustment strategy to gain better control over the optimization process. Specifically, we resumed training from the checkpoint saved at epoch 3 and reduced the learning rate to 1e-6. This two-stage fine-tuning approach led to a significant improvement in F1-score, which is shown in Table 3 and ultimately helped us achieve top-rank in the competition.

Epoch	Validation Loss	F1
1	**0.7025**	**0.7129**
2	0.7256	0.7116

Table 3
After two-stage fine-tuning (resumed from epoch 3)

However, this fine-tuning approach was conceptualized during the final phase of development, leaving insufficient time to implement and evaluate it in the remaining tasks, namely binary English, binary Spanish and multiclass Spanish. Exploring the impact of this strategy on these tasks remains a promising direction for future research.

4. Error Analysis



(a) Confusion matrix for binary classification of English texts.

(b) Confusion matrix for multiclass classification of English texts.

Figure 2: Confusion matrix for Hope Classification of English texts.

4.1. Binary English Classifications

Figure 2a illustrates the confusion matrix for the binary English classification task, where the two classes are *Hope* and *Not Hope*. The model correctly predicted 956 out of 1003 *Not Hope* instances and 868 out of 899 *Hope* instances, demonstrating strong classification performance.

There were 47 false positives (predicting *Hope* when the true label is *Not Hope*) and 31 false negatives (predicting *Not Hope* when the true label is *Hope*). These low mistaken classification rates indicate a well-balanced model with high precision and recall across both classes, especially in detecting hopeful content.

Table 4
Examples of Binary English Mispredicted Samples

Text (Preprocessed)	Actual Label	Predicted Label
hopeful that scotus ends its term on wednesday w the issuance of 4 remaining opinions i really need a nap	Not Hope	Hope
hope is a good thing maybe the best thing and no good thing ever dies stephenking	Not Hope	Hope
woke people accuse conservatives of narrowminded thinking but theyre the ones with a narrowminded perspective hence their persistent desire to change everything when you effect the change they yearn for they begin to see the bigger picture and call for some things	Hope	Not Hope
im having a bad time rn and i feel bad and i am bad and everything is bad and i yearn for but yet	Not Hope	Hope
no its not pessimistic to anticipate contingencies be prepared take appropriate precautions and avoid foolhardy risks	Hope	Not Hope

Despite the overall strong performance, several misclassified cases reveal specific challenges in hope detection, which are represented in Table 4, implying its limitations.

- **Lexical cues vs. intent (e.g., "hopeful that scotus ends its term..."))**: Although the word "hopeful" is present, the context is procedural and factual. The model likely relied on surface-level lexical cues without recognizing the absence of genuine emotional or aspirational intent.
- **Famous quotes or abstract generalizations (e.g., "hope is a good thing maybe the best thing..."))** : This quote, from *The Shawshank Redemption*, reflects a general life philosophy rather than a personal or contextualized expression of hope. While the model labeled it as *Hope* based on sentiment, it doesn't align with the annotation guidelines.
- **Ambiguity and incomplete emotional expressions (e.g., "i yearn for but yet..."))** : This phrase conveys a sense of longing but lacks clarity or specificity. Such vague language makes it difficult for the model to interpret correctly, often resulting in a false positive.
- **Complex reasoning or critical rhetoric (e.g., "woke people accuse conservatives of narrowmind..."))**: Although labeled as *Hope*, this statement functions as a political critique with an implicit sense of optimism for change. The absence of explicit hopeful language may have caused the model to misclassify it as *Not Hope*.

4.2. Multiclass English Classifications

Figure 2b presents the confusion matrix for the multiclass English classification task, which includes five categories: *Generalized Hope*, *Not Hope*, *Realistic Hope*, *Sarcasm*, and *Unrealistic Hope*.

The model performs best on the *Not Hope* and *Sarcasm* classes, with 763 and 242 correct predictions respectively, showing that it can clearly identify non-hopeful and sarcastic content. *Generalized Hope* and *Realistic Hope* are also handled reasonably well, even though some confusion between them suggests overlapping language patterns. Meanwhile, the *Unrealistic Hope* class remains the most difficult to classify accurately, probably because of its subtle distinctions from the other hopeful categories.

Table 5
Examples of Multiclass English Mispredicted Samples

Text (Preprocessed)	Actual Label	Predicted Label
ik it isn't gonna be there but I hope there will be leviathan axe in the last section	Generalized Hope	Realistic Hope
this is awful please pray for these poor people no one should have died that way but will this administration do anything nope they have a clown tribunal to attend to and a constitution to ignore	Generalized Hope	Not Hope
in other words they anticipate home values to crater in canada in the coming 1216 months and will deploy capital opportunistically in canada	Realistic Hope	Not Hope
i am starting to think that i distance myself from my emotions that is all of them when i anticipate feeling a negative one	Generalized Hope	Not Hope
25000 only because im not rich or well off so 25gs is a lot id rather have the guarantee for small life time changes than twiddling my thumbs hoping for great luck on one bet	Unrealistic Hope	Not Hope

Table 5 shows several misclassified examples from the multiclass English classification task. These samples reveals the key challenges our model faced in detecting fine-grained categories of hope as emotion:

- **Generalized vs. Realistic Hope (e.g., "I hope there will be leviathan axe...")** This example was labeled as *Generalized Hope* due to its broad and wishful nature. However, the model predicted

Realistic Hope, likely influenced by the mention of a specific object ("leviathan axe"). This suggests that the presence of concrete terms can bias the model toward interpreting hope as realistic, even when the expectation itself is not well grounded.

- **Implicit hopeful intent masked by distress (e.g., "please pray for these poor people...")** The statement conveys emotional concern and a plea for support, aligning with *Generalized Hope*. Yet the model classified it as *Not Hope*, likely due to its distressing content and negative tone. This reflects the model's difficulty in recognizing subtle expressions of hope when embedded in emotionally heavy or tragic contexts.
- **Domain-specific optimism not recognized (e.g., "anticipate home values to crater...")** This sentence includes a market prediction followed by an implicitly optimistic investment stance, labeled as *Realistic Hope*. However, the model predicted *Not Hope*, possibly due to unfamiliarity with financial language or the sentence's neutral tone, despite its forward-looking intent.
- **Psychological self-reflection lacking explicit optimism (e.g., "i distance myself from my emotions...")** Although this introspective statement was labeled as *Generalized Hope*, it lacks overt positivity. The model predicted *Not Hope*, suggesting difficulty in interpreting abstract or internal expressions of hope when they are not explicitly stated.
- **Unrealistic expectations presented rationally (e.g., "25gs is a lot... hope for great luck in one bet")** This sentence describes an unlikely but hopeful situation, correctly labeled as *Unrealistic Hope*. Yet the model predicted *Not Hope*, likely influenced by the rational tone and comparative phrasing. This illustrates the challenge of detecting hope when it is framed logically rather than emotionally.

These examples highlight common limitations in both binary and multiclass hope classification, especially in handling implicit expressions, overlapping category boundaries, and contextual ambiguity. Enhancing model performance may require more sophisticated semantic modeling, greater sensitivity to tone and discourse style, and targeted data augmentation to better represent underexplored types of hope.

4.3. Explanation

The high performance scores can be attributed not only to the use of a combined training set—formed by concatenating the original training and development datasets, as noted in Section 3.1.2, but also to the strong computational capabilities of the Transformer architecture. By increasing the amount of training data, the model gains exposure to more diverse psychological and emotional patterns, allowing it to learn better representations without requiring further fine-tuning. Furthermore, the self-attention mechanism and parallel computation inherent in our models enable efficient learning of complex dependencies. As a result, the evaluation metrics are notably high, which is expected given both the enriched training data.

5. Main Results

The official submission results and full leaderboards of each subtask are shown in Table 6, 7, 8, 9. In terms of the Weighted F1 score, we obtained 0.8611 on Subtask 1a: English Binary Classification, ranked 11th place. We achieved a surprisingly strong result of 0.7851 on Subtask 2a: English Multiclass Classification, preserving the first place. Additionally, we obtained 0.8377 on Subtask 2a: Spanish Binary Classification, ranked the 5th place. Finally, our best result on Subtask 2b: Spanish Multiclass Classification, ranked third place, is lower than the Weighted F1-score of the Top 1 and Top 2 teams, which are -0.0208 and -0.0109, in turn.

Pos.	Team Name	Macro F1	Weighted F1
1	michaelibrahim	0.8713	0.8719
2	rogeliorjr1	0.8704	0.8707
3	nayeem01	0.8701	0.8707
4	…	…	…
11	**supachoke (Us)**	**0.8608**	**0.8611**
12	ebuka	0.8597	0.8597
13	tafredri	0.8597	0.8597

Table 6
Official Results for Subtask 1a: English Binary Classification

Pos.	Username	Macro F1	Weighted F1
1	**supachoke (Us)**	**0.7546**	**0.7851**
2	ebuka	0.7484	0.7881
3	tafredri	0.7484	0.7881
4	lephuquy	0.7425	0.7838
5	michaelibrahim	0.7420	0.7819

Table 7
Official Results for Subtask 1b: English Multiclass Classification

Pos.	Team Name	Macro F1	Weighted F1
1	teddymas	0.8521	0.8520
2	abit7431	0.8464	0.8463
3	lephuquy	0.8446	0.8451
4	dmadera	0.8435	0.8434
5	**supachoke (Us)**	**0.8376**	**0.8377**
6	nayeem01	0.8247	0.8252
7	michaelibrahim	0.7896	0.7902

Table 8
Official Results for Subtask 2a: Spanish Binary Classification

Pos.	Team Name	Macro F1	Weighted F1
1	lephuquy	0.7417	0.7682
2	dmadera	0.7221	0.7583
3	**supachoke (Us)**	**0.6984**	**0.7474**
4	teddymas	0.6901	0.7079
5	nayeem01	0.6795	0.7080
6	michaelibrahim	0.6545	0.7002

Table 9
Official Results for Subtask 2b: Spanish Multiclass Classification

Note: All scores are rounded to the fourth decimal place.

6. Conclusion

In this study, we share our team's contributions to the PolyHope shared task at IberLEF 2025, which focused on detecting hope speech in multilingual social media texts. Our approach was based on leveraging pre-trained BERT-based models, particularly the BGE-large-en-v1.5 and BGE-m3, which performed strongly across both binary and multiclass classification tasks for English and Spanish.

Through a series of experiments, we explored different strategies, our study reveals key insights, besides fine-tuning methods, when using data concatenation, our models could easily learn on boarder context, improving performance of hope detection, but still be limited on detecting more complex emotional patterns of hope. These methods helped us achieve competitive results, including 11th place

in English binary classification, 1st place in English multiclass, 5th in Spanish binary, and 3rd in Spanish multiclass classification. These rankings highlight the effectiveness of our multilingual approach. One key factor behind these results was the use of a two-stage fine-tuning process, which led to noticeable performance gains. While the outcomes were promising, there's still room to improve, especially in refining preprocessing steps and further exploring fine-tuning techniques across tasks. Future work could explore ensemble approaches combining the strengths of different architectures, and contextual extraction, cross-cultural exploration for hope expressions classifiers. Additionally, further investigation into the impact of preprocessing strategies could help explain the performance differences between our original and extended implementations. The advancement of computational methods for hope detection enables novel applications in mental health monitoring, social media analysis, and discourse studies. We believe these improvements could boost the adaptability and accuracy of hope speech detection systems in the future. Overall, our findings underscore the strong potential of transformer-based models in handling nuanced sentiment classification and their subcategories, particularly in the challenging area of hope speech detection.

Acknowledgments

We would like to express our deepest appreciation to **Nguyen Tien Thang** for his unwavering support and expert mentorship during the model training process, which significantly improved the quality of our results. Our sincere thanks also go to **Le Duc Tai** for his critical review and insightful suggestions that greatly enhanced the clarity and rigor of this manuscript. Finally, we are profoundly grateful to **Dr. Nguyen Trong Chinh**, my esteemed supervisor, whose scholarly guidance, constructive feedback, and steadfast encouragement have been instrumental throughout this research journey.

Declaration on Generative AI

The authors acknowledge limited use of generative AI tools for grammar and phrasing. No AI systems were used for data collection, analysis, or interpretation. All final content was reviewed and approved by the authors.

References

[1] 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.

[2] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078.

[3] 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.

[4] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. A. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, D. García-Baena, J. A. García-Díaz, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion organized as part of ACL 2022, 2022, pp. 378–388. doi:10.18653/v1/2022.ltedi-1.58.

[5] D. García-Baena, F. Balouchzahi, S. Butt, M. A. García-Cumbreras, A. L. Tonja, J. A. García-Díaz, 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, Procesamiento del Lenguaje Natural 73 (2024) 407–419.

[6] S. M. Jiménez-Zafra, M. A. García-Cumbreras, D. García-Baena, J. A. García-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. U. na López, Overview of hope at iberlef 2023: Multilingual hope speech detection, in: Procesamiento del Lenguaje Natural, volume 71, 2023, pp. 371–381.
[7] C. Snyder, Hypothesis: There is hope, in: Handbook of Hope, Elsevier, 2000, pp. 3–21.
[8] C. Snyder, B. Hoza, W. E. Pelham, M. Rapoff, L. Ware, M. Danovsky, L. Highberger, H. Ribinstein, K. Stahl, The development and validation of the children's hope scale, Journal of Pediatric Psychology 22 (1997) 399–421.
[9] S. Butt, F. Balouchzahi, M. Amjad, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview of polyhope at iberlef 2025: Optimism, expectation or sarcasm?, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[10] S. Butt, F. Balouchzahi, A. I. Amjad, M. Amjad, H. G. Ceballos, S. M. Jiménez-Zafra, Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english, https://doi.org/10.13140/RG.2.2.19761.90724, 2025. Preprint on ResearchGate.
[11] BAAI, bge-m3, https://huggingface.co/BAAI/bge-m3, 2024. Accessed: 2025-04-10.
[12] BAAI, bge-large-en-v1.5, https://huggingface.co/BAAI/bge-large-en-v1.5, 2023. Accessed: 2025-04-10.
[13] C. Snyder, The Psychology of Hope: You Can Get There from Here, Simon and Schuster, 1994.
[14] C. Snyder, Hope theory: Rainbows in the mind, Psychological Inquiry 13 (2002) 249–275.
[15] 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) 3983. doi:10.3390/app13063983.
[16] M. Krasitskii, O. Kolesnikova, L. C. Hernandez, G. Sidorov, A. Gelbukh, HOPE2024@IberLEF: A cross-linguistic exploration of hope speech detection in social media, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), CEUR Workshop Proceedings, 2024. URL: https://www.researchgate.net/publication/385383738_HOPE2024IberLEF_A_Cross-Linguistic_Exploration_of_Hope_Speech_Detection_in_Social_Media, cEUR-WS.org, Vol. 3756.
[17] R. Pan, Ángela Almela, G. Alcaraz-Mármol, Umuteam at hope@iberlef 2024: Fine-tuning approach with sentiment and emotion features for hope speech detection, in: Proceedings of the HOPE Workshop at IberLEF 2024, volume 3756, CEUR-WS.org, 2024, pp. 9–16. URL: https://ceur-ws.org/Vol-3756/HOPE2024_paper14.pdf.
[18] B. R. Chakravarthi, 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, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.
[19] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, J. W. Kim, Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism, Applied Sciences 10 (2020) 5841. URL: https://doi.org/10.3390/app10175841. doi:10.3390/app10175841.
[20] B. Athiwaratkun, A. G. Wilson, A. Anandkumar, Probabilistic fasttext for multi-sense word embeddings, arXiv preprint arXiv:1806.02901 (2018). URL: https://arxiv.org/abs/1806.02901.
[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[22] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020). URL: https://arxiv.org/abs/2006.03654.
[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.
[24] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL: https://arxiv.org/abs/2402.03216. arXiv:2402.03216.
[25] C. Xu, Y. Shen, Y. Zhang, Q. Zhang, Y. Zhang, Z. Liu, M. Sun, Baai general embedding (bge): A new embedding model family for retrieval and beyond, arXiv preprint arXiv:2309.16609 (2023).

URL: https://arxiv.org/abs/2309.16609.
[26] Microsoft, microsoft/deberta-v3-large, https://huggingface.co/microsoft/deberta-v3-large, 2021. Accessed: 2025-04-22.
[27] F. AI, Facebookai/roberta-large, https://huggingface.co/FacebookAI/roberta-large, 2019. Accessed: 2025-04-22.