LPQ Team at HOPE 2025: Multilingual Hope Speech Detection Using BERTology and LLM Adaptation

Le Phu Quy[1,2,*], Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents our approach for detecting hope speech in social media posts written in English and Spanish. The system employs integrated transformer models, while separate experiments with the LLaMA model are conducted for emotion detection. Based on the PolyHope task at IberLEF 2025, the framework supports both binary classification—separating hope speech from non-hope speech—and a detailed multiclass categorization that distinguishes among generalized, realistic, and unrealistic expressions of hope. Experimental results show strong performance across different models and languages.

Keywords

Hope classification, Spanish language, English language, sentiment analysis, fine-tuning BERT

1. Introduction

Hope is an essential human emotion that shapes behavior, mood, and decision-making across diverse cultural contexts, and it fills our lives with both resilience and aspiration. In today's digital era, social media platforms have become vibrant arenas where people share intimate personal dreams alongside collective visions for a brighter future. While the academic challenge of detecting and analyzing hope speech is made more complex by its subtle blend of irony, sarcasm, and multilingual variations, developing robust methodologies to capture these nuances remains crucial. Not only does such work contribute to our scholarly understanding of emotional expression, but it also supports practical efforts in mental health monitoring, crisis intervention, and designing public policies that truly resonate with human experience.

The PolyHope shared task at IberLEF 2025 [1] aims to detect and classify hope speech in social media posts written in English and Spanish. This effort is significant in the field of computing and sentiment analysis because it attempts to differentiate various forms of hope instead of limiting captures as a mere positive or negative classification. The two main tasks in this challenge are: (1) hope speech detection using binary classification which involves hopeful messages and non-hopeful messages, and (2) multiclass differentiation of general, realistic, and unrealistic hope alongside sarcastic remarks.

In this work, we present our system for detecting hope speech in social media posts across both English and Spanish. By integrating leading transformer models—DeBERTa and DistilBERT for English, and BETO alongside RoBERTa-BNE for Spanish—while also using LLaMA for emotion detection, our system addresses both binary and multiclass classification (capturing generalized, realistic, unrealistic, and sarcastic forms of hope). Extensive experiments on the PolyHope 2025 dataset not only confirm the robustness of our approach but also pave the way for practical applications in mental health monitoring, crisis response, and political discourse analysis.

IberLEF 2025, September 2025, Zaragoza, Spain

[*]Corresponding author.

23521316@gm.uit.edu.vn (L. P. Quy); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0000-0001-8340-1405 (D. V. Thin)

2. Related Work

Transformer-based models have greatly changed how we handle many NLP tasks, including emotion and sentiment analysis. For example, BERT [2] and its variants perform very well on sentiment benchmarks. For multilingual tasks, models like XLM-R [3] and mBERT [4] let us transfer learning from languages with lots of data to those with less data. In hope speech detection, transformer approaches have also shown promising results. At IberLEF 2023, the I2C-Huelva team [5] reached the top rankings by using BERTuit for Spanish and BERT for English. Similarly, NLP URJC [6] used BERT for English and BETO for Spanish. These cases support the trend that transformer-based models generally beat traditional machine learning methods in emotion detection [7].

Language-specific transformer models are very effective for emotion tasks. DeBERTa [8] uses a special attention mechanism and an improved mask decoder, leading to state-of-the-art performance on English sentiment analysis. For Spanish, BETO [9] and RoBERTa-BNE [10] are strong choices; notably, RoBERTa-BNE was trained on a massive corpus of Spanish text from the National Library of Spain. Large language models (LLMs) such as GPT-3 [11] and LLaMA [12] have boosted emotion and sentiment analysis by understanding subtle language cues. For example, the Zootopi team at IberLEF 2023 used prompt-based ChatGPT to detect hope speech, winning the Spanish subtask [13]. To reduce the high computational cost of fine-tuning huge models, researchers now use Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) [14].

3. Methodology

3.1. Data Preprocessing

In the early stages of our research, we experimented with a variety of preprocessing techniques—including text normalization, comprehensive punctuation removal, and aggressive stop-word filtering—in an effort to enhance our model's performance in detecting hope speech. However, we found that these approaches often eliminated important contextual and emotional nuances that are crucial for understanding the subtleties of human communication. Based on our experimental insights, we ultimately adopted a minimal preprocessing strategy. This approach relies on standard, model-specific tokenization and essential sequence management, ensuring that the delicate linguistic details remain intact. Such a strategy not only preserves the richness of the original text but also supports our aim of achieving robust and sensitive emotion detection in both English and Spanish social media content.

3.2. BERT-Base Classification

Recent studies have demonstrated the effectiveness of BERT-based approaches for text classification tasks. All models are fine-tuned within a unified framework to optimize performance on our hope speech classification task. For the BERT-based models, we minimize a standard cross-entropy loss using hyperparameters recommended by recent literature and validated through empirical studies. This BERT-based classification framework effectively combines deep and contextualized language representations with task-specific fine-tuning, making it well-suited for capturing subtle cues in text. Its widespread adoption in recent research underscores its robustness and effectiveness across diverse text classification settings.

The following models were incorporated in our experiments:

- **DeBERTa**: Utilizes a disentangled attention mechanism to enhance contextual representation.
- **DistilBERT**: A lighter version of BERT, retaining approximately 97% of its performance while significantly reducing computational overhead.
- **BETO**: A Spanish-specific adaptation of the BERT architecture, optimized to capture native linguistic nuances.
- **RoBERTa-BNE**: A RoBERTa variant fine-tuned on extensive Spanish corpora, enabling effective handling of regional dialects and stylistic variations.

3.3. LLaMA Hybrid Approach

Large language models perform well in many NLP tasks, yet adapting them for specialized domains like hope speech classification requires balancing efficiency with domain-specific adjustments. The LLaMA hybrid approach achieves this by combining parameter-efficient adaptation and structured instruction tuning, preserving the model's contextual depth while enhancing its sensitivity to subtle signals of hope. The framework follows a four-stage process: Low-Rank Adaptation, structured prompt engineering, targeted instruction tuning, and post-processing.

1. **LoRA Configuration**: Low-Rank Adaptation (LoRA) is employed by decomposing weight updates into low-rank matrices. This strategy enables fine-tuning of the LLaMA model by adjusting only a small subset of its parameters, thereby preserving pre-trained performance while optimizing computational efficiency. For these experiments, a rank of $r = 16$ and a scaling factor $\alpha = 32$ were selected, providing an optimal trade-off between adaptation capacity and resource demands.
2. **Structured Prompt Engineering**: The task is rephrased as a guided conversation with clear instructions to help the model detect hope. Detailed guidelines help the model identify clear signs of hope, subtle hints, and the intended meaning behind sarcastic remarks. This organized method ensures the model stays sensitive to the fine nuances of human communication.
3. **Instruction Tuning**: The SFTTrainer uses a response-focused loss function to increase the likelihood of correct labels. This tuning method takes into account previously generated tokens and the parameters adjusted by LoRA, fine-tuning the model for accurate hope speech classification while keeping computations efficient.
4. **LLaMA Post-Processing**: A dedicated post-processing module extracts and refines predictions from the generated text. Using multiple extraction strategies, it effectively handles variations in output formats, ensuring consistent and reliable classification results.

4. Experiments

4.1. Dataset

The dataset [15, 16, 17, 18, 19, 20, 21] consists of English (en) and Spanish (es) social media texts annotated for two tasks: binary classification (*Hope* vs. *Not Hope*) and multiclass classification. Statistics are summarized below:

Language	Split	Samples	Total	Not Hope (%)	Hope (%)
English	Train	5,233	7,135	53.6	46.4
	Dev	1,902		52.7	47.3
Spanish	Train	11,243	15,331	52.7	47.3
	Dev	4,088		52.9	47.1

Table 1: Combined Dataset Statistics and Binary Class Distribution by Language

Language	Not Hope	Generalized Hope	Realistic Hope	Unrealistic Hope	Sarcasm
English (Train)	42.9%	24.5%	10.3%	9.0%	13.2%
English (Dev)	42.9%	24.5%	10.3%	9.0%	13.2%
Spanish (Train)	47.9%	24.5%	9.9%	11.6%	6.2%
Spanish (Dev)	47.9%	24.5%	9.9%	11.6%	6.2%

Table 2: Multiclass Label Distribution

4.2. Experiment Setting

The experimental setup aimed to assess model performance in classifying English and Spanish social media texts expressing hope, across both binary and multiclass tasks. To ensure a balanced comparison, four transformer-based architectures were tested. DeBERTa (microsoft/deberta-base) stood out for its advanced attention mechanism, which improves understanding of complex language patterns. DistilBERT (distilbert-base-uncased) offered a faster, lightweight alternative to BERT, maintaining 97% of its performance with fewer computational demands. RoBERTa-BNE (BSC-TeMU/roberta-base-bne) was specialized for Spanish, leveraging data from the National Library of Spain to refine its understanding of the language. Similarly, BETO (dccuchile/bert-base-spanish-wwm-cased) adapted BERT's framework for Spanish, using whole-word masking to better capture meaning.
Consistency in training was a priority across all experiments. Each model was fine-tuned with a batch size of 8, using the AdamW optimizer at a learning rate of 2×10^{-5} and a weight decay of 0.01. Models were trained for 5 epochs with early stopping, and input text was limited to 128 tokens through truncation or padding. These parameters adhered to best practices for transformer models while also ensuring computational efficiency.
For the Llama-3 experiments (`unsloth/Meta-Llama-3.1-8B-bnb-4bit`), adjustments were made to address resource limitations. Techniques such as 4-bit quantization, implemented via `bitsandbytes`, reduced GPU memory usage by 70%. Low-Rank Adaptation (LoRA) was applied with a rank of $r = 16$ and scaling factor $\alpha = 32$, allowing efficient fine-tuning by updating only 0.1% of the model's parameters. Prompts were tailored to the task: binary classification relied on simple instructions focused on detecting hope versus no hope, while multiclass classification required more detailed prompts to distinguish between nuanced categories, including generalized, realistic, unrealistic hope, and sarcasm. Training for Llama-3 utilized an extended context window of 512 tokens, a higher learning rate of 2×10^{-4}, and a shorter training cycle of 3 epochs to manage computational constraints. This experimental design ensured a robust evaluation, comparing traditional transformer-based architectures against the advanced capabilities of Llama-3 while maintaining efficiency and academic rigor

4.3. Main Result

DeBERTa achieves state-of-the-art performance in both binary (84.76% F1) and multiclass (75.83% F1) hope speech detection, outperforming DistilBERT and Llama 3.1 8B across languages. While transformer models generalize robustly, multiclass tasks remain challenging due to nuanced hope subtypes and sarcasm. Llama 3.1 8B shows promise but lags in multiclass scenarios (72.56% F1), suggesting room for improved fine-tuning strategies. Results validate the efficacy of task-specific architectures over LLM approaches.

Our approach demonstrates strong performance across a diverse range of classification tasks. In particular, our model achieved first rank in Multiclass Spanish, recording a weighted F1 score of 0.7417, and secured third place in Spanish Binary with an F1 score of 0.8446. These results highlight the robustness of language-specific tuning in Spanish contexts. However, for English tasks, our model performs worse on English tasks, ranking seventh in Binary and fourth in Multiclass. This lower performance may be due to the greater complexity of the English dataset and less effective fine-tuning. Detailed evaluation shows that transformer models like DeBERTa and BETO generally outperform the Llama 3.1 8B setup by approximately 2–4% F1.

5. Error Analysis and Discussion

Our analysis revealed two primary error patterns affecting model performance. First, the models struggled to accurately distinguish between subtle hope subtypes, particularly in Spanish texts. Differentiating between Realistic Hope and Unrealistic Hope proved especially difficult, as reflected in F1-scores ranging from 0.51 to 0.56. Second, data imbalance played a significant role in these challenges. The Sarcasm class accounts for only 6.2% of the Spanish training data and exhibited low recall, between

Subtask	Model	Accuracy	Precision	Recall	F1 Score
Binary Classification					
1.a English	DeBERTa	84.75%	84.80%	84.70%	84.76%
	DistilBERT	84.02%	84.10%	83.90%	84.02%
	Llama 3.1 8B	83.28%	83.40%	83.10%	83.30%
1.b Spanish	BETO	83.90%	83.80%	84.00%	83.92%
	RoBERTa-BNE	83.98%	84.00%	83.90%	83.99%
	Llama 3.1 8B	83.20%	83.30%	83.10%	83.25%
Multiclass Classification					
2.a English	DeBERTa	75.55%	75.60%	76.00%	75.83%
	DistilBERT	74.55%	74.60%	74.80%	74.71%
	Llama 3.1 8B	72.55%	72.70%	72.40%	72.56%
2.b Spanish	BETO	75.56%	76.10%	76.00%	76.08%
	RoBERTa-BNE	74.85%	75.00%	74.80%	74.90%
	Llama 3.1 8B	72.46%	72.60%	72.40%	72.50%

Table 3: Experimental Results of Binary and Multiclass Classification

Task	Rank	Team Name	Accuracy	Avg Macro F1
English Binary	1	michaelibrahim	0.8718	0.8713
	2	rogeliorjr1	0.8705	0.8703
	7	**lephuquy (Ours)**	**0.8663**	**0.8659**
English Multiclass	1	supachoke	0.7815	0.7546
	2	ebuka	0.7903	0.7484
	4	**lephuquy (Ours)**	**0.7878**	**0.7425**
Spanish Binary	1	teddymas	0.8520	0.8520
	2	abit7431	0.8464	0.8463
	3	**lephuquy (Ours)**	**0.8450**	**0.8445**
Spanish Multiclass	2	dmadera	0.7540	0.7221
	3	supachoke	0.7474	0.6983
	1	**lephuquy (Ours)**	**0.7677**	**0.7416**

Table 4: PolyHope at IberLEF Final Ranking

0.60 and 0.77, because the models tended to favor the majority Not Hope class, which made up 47.9% of the data. These findings highlight the importance of applying targeted sampling strategies to improve minority class representation and, ultimately, enhance overall model performance in detecting nuanced expressions of hope.

While our methodology shows impressive baseline performance, there remain several notable limitations that must be discussed. For instance, the current prompting strategies fall short in capturing subtle sarcasm and context-dependent expressions of hope. To overcome these issues, we propose three avenues for future exploration. First, implementing chain-of-thought prompting might allow models to break down complex utterances into distinct contextual signals prior to making a final decision. Second, employing an ensemble of transformer models—by combining predictions from architectures such as DeBERTa, BETO, and Llama—could leverage their unique strengths, especially in low-resource scenarios for sarcasm detection. Finally, integrating culture-aware data augmentation techniques may help reduce dialectical biases in Spanish hope speech by utilizing region-specific lexicons to boost generalization. These limitations underscore the need for further research into improved fine-tuning methods and innovative data augmentation practices to elevate overall performance.

5.1. Conclusion

In summary, the system showed robust performance in detecting hope speech on English and Spanish social media posts, with binary F1 scores around 0.83–0.84 achieved by transformer-based models. However, accurately distinguishing the subtle differences among hope expressions remains a challenge, especially when implicit signals mix with sarcasm. In our experiments, the LLaMA model was also employed for emotion detection. While it demonstrated promising potential, its performance has not yet surpassed that of the established transformer architectures. This indicates that despite LLaMA's advanced capabilities, further refinement in instruction tuning and architectural adjustments is needed for it to effectively capture subtle linguistic signals in hope speech detection. Future work should aim to train models that understand context, adjust to different dialects, and combine the strengths of both transformer and LLaMA methods to boost accuracy.

Acknowledgements

This research was supported by the VNUHCM-University of Information Technology's Scientific Research Support Fund.

Declaration on Generative AI

Generative AI tools were employed only to enhance linguistic clarity. All data, analyses, and conclusions were conceived, executed, and validated by the authors. The authors retain full accountability for every aspect of the manuscript's content.

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] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the NAACL, 2019, pp. 4171–4186.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[4] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint arXiv:1906.01502 (2019).

[5] J. L. D. Olmedo, J. M. Vázquez, V. P. Álvarez, I2c-huelva at hope2023@ iberlef: Simple use of transformers for automatic hope speech detection (2023).

[6] M. Á. Rodríguez-García, A. Riaño-Martínez, S. Montalvo-Herranz, Urjc-team at hope2023@ iberlef: Multilingual hope speech detection using transformers architecture., in: IberLEF@ SEPLN, 2023.

[7] S. J. Lee, J. Lim, L. Paas, H. S. Ahn, Transformer transfer learning emotion detection model: synchronizing socially agreed and self-reported emotions in big data, Neural Computing and Applications 35 (2023) 10945–10956.

[8] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).

[10] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Spanish language models, arXiv preprint arXiv:2107.07253 (2021).

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[13] A. Ngo, H. T. H. Tran, Zootopi at hope2023@ iberlef: Is zero-shot chatgpt the future of hope speech detection?, in: IberLEF@ SEPLN, 2023.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

[15] 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?, Procesamiento del Lenguaje Natural (2025).

[16] 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, 2025.

[17] G. Sidorov, F. Balouchzahi, L. Ramos, H. Gómez-Adorno, A. Gelbukh, MIND-HOPE: Multilingual Identification of Nuanced Dimensions of HOPE, Technical Report, 2024.

[18] 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.

[19] 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.

[20] D. García-Baena, F. Balouchzahi, S. Butt, M. 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, Procesamiento del Lenguaje Natural 73 (2024) 407–419.

[21] D. García-Baena, M. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, V. G. Rafael, Hope speech detection in spanish: The LGTB case, Language Resources and Evaluation (2023) 1–31.