Choosing the Right Language Model for the Right Task

Chau Pham Quoc Hung[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

In recent years, many language models have been introduced to tackle various NLP tasks. In general, most pre-trained language models achieve superior performance to traditional machine learning algorithms. However, with different architectures and pre-training corpora, some models are proven to be more efficient than others in some specific domains. By participating in the IberLEF 2024 HOPE shared tasks, we aim to explore the effectiveness of pre-trained language models in identifying hope speech in text by fine-tuning different models and compare their performance for each given task. This simple yet effective approach helps us achieve satisfactory results in the competition, with all tasks getting 3rd rank or higher.

Keywords

Hope Speech, Pre-trained Language Model, Transformer, Hope Classification, Spanish Language, English Language

1. Introduction

Hope has been widely recognized as a crucial and inherent aspect of human existence [1], a feeling that is both powerful and complex. Serving as an existential element within each individual, hope manifests itself in various ways in both private and public spheres [2]. It is essentially a positive outlook on things, a belief that good things are possible [3].

Hope serves as a motivational force, fostering the belief that one's actions can influence the desired future state. It fuels the perception that effort has the potential to lead to positive change. In essence, hope is a complex psychological construct that motivates individuals to strive for positive outcomes in the face of uncertainty [4].

Due to its relevancy, especially in the modern era, a need to detect hope speech in natural language is becoming more prominent. However, manually carrying out such a task is laborious and costly. Thus, many machine-learning systems have been employed for this specific task. In that sense, The HOPE shared task [5], at IberLEF 2024 [6], challenges its participants to design an automatic system to detect hope speech in 2 tasks:

- **Task 1 - Hope for Equality, Diversity and Inclusion**: This task aims to detect hope speech, focusing on promoting the inclusion of vulnerable groups and ultimately achieving Equality, Diversity, and Inclusion. The main objective is to identify whether a given Spanish tweet contains hope speech or not.
- **Task 2 - Hope as expectations**: This task focuses on expectations, desirable and undesirable facts. There are two subtasks in this task, specifiacally: Binary Hope Speech Detection and Multiclass Hope Speech Detection. Both subtasks are required for English and Spanish text.

After our participation, we present a summary of the work and the results obtained as shown in this paper, including preprocessing input text, fine-tuning and comparing multiple pre-trained language models for both Spanish and English to draw out the conclusion: *Which model fits which task?*

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
20521360@gm.uit.edu.vn (C. P. Q. Hung); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0009-0001-6497-204X (C. P. Q. Hung); 0000-0001-8340-1405 (D. V. Thin)

The structure of this paper is as follows: In Section 2, we describe some previous works on hope speech detection and comment on their performance. In Section 3, we detail our methodology for detecting hope speech in the assigned tasks. In Section 4, we describe the datasets provided by the organizers. In Section 5, we experiment and evaluate our proposed methods to see which one is the most suitable. In Section 6, we apply our best method to the unlabeled datasets and explain the results. Finally, in Section 6, we conclude the paper and speculate possible improvements in the future.

2. Related Work

Pioneering the field of hope speech detection on social media platforms by aligning hope with equality, diversity, and inclusion (EDI), Chakravarthi [4] created the HopeEDI corpus using YouTube comments in both Dravidian and English languages, including English as well as code-mixed Tamil-English and Malayalam-English datasets. This led to the shared task at the First Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-EACL) [2] in 2021. Later it was expanded to include Spanish and code-mixed Kannada-English texts and used in the second next edition shared task at the same workshop in 2022 [7]. Experimenting with various Machine Learning models and a CNN model for detecting hope speech. Chakravarthi concluded the proposed CNN model achieves the highest F1-score across all languages [8].

In the LT-EDI-EACL 2021 workshop, Dowlagar and Mamidi [9] used 3 models for the task: SVM as baseline, multilingual BERT, and multilingual BERT with CNN classifier. The latter model outperformed the others. Arunima et al. [10] fine-tuned mBERT for Malayalam and Tamil then used BERT for English. This simple yet effective approach yielded competitive results. Upadhyay et al. [11] tried two approaches: contextual embeddings with classifiers and majority voting ensemble by fine-tuning pre-trained transformer models (BERT [12], ALBERT [13], RoBERTa [14], IndicBERT [15]).

At shared task Multilingual Hope Speech detection (HOPE) of IberLEF 2023 [16], Zahra et al. [17] employed SVM for English dataset and KNN for Spanish dataset. The approach achieved third place in both datasets. Inspired by previous works in hope speech detection, Moein et al [18] used CNN and achieved fourth place in both sub-tasks. Juan et al. [19] fine-tuned separate transformer-based models for different languages: DistilBERT [20] for English which got first rank and BERTuit [21] for Spanish which got second rank.

From a performance standpoint, models that integrate pre-trained language models have proven to be highly successful in hope speech detection and yielded great results as demonstrated in various studies cited above.

3. Methodology

Inspired by the success of pre-trained language models in recent years, we employed multiple models for the classification problems in both tasks. These pre-trained models are based on the transformer [22] architecture and can be fine-tuned for specific problems. We used both monolingual and multilingual models and then evaluated the performance of each model to conclude which model suits which task. We used BERT[12], DistilBERT [20], RoBERTa [14] for English, BETO [23], SpanBERTa [1], RoBERTuito [24] for Spanish, and mBERT [2], XLM-R [25] for both languages. All of the models were applied by using the *Transformer* library [3] provided by *HuggingFace* [4]. The summarized details of the used pre-trained language models are shown in Table 1.

To apply the mentioned Pre-trained language models for the classification problem in 2 tasks, we added a classification layer on top of the output for the [CLS] token, which represents sentence-level classification. The number of outputs in the classification layer depends on the number of classes for

[1] https://github.com/chriskhanhtran/spanish-bert
[2] https://github.com/google-research/bert/blob/master/multilingual.md
[3] https://github.com/huggingface/transformers
[4] https://huggingface.co/

Table 1
Summarized information of various pre-trained language models used in our approach

Model	Pre-trained data	Tokenization	HuggingFace model's name
English language models			
BERT	BooksCorpus + English Wikipedia	WordPiece	google-bert/ bert-base-uncased
DistilBERT	BooksCorpus + English Wikipedia	WordPiece	distilbert/ distilbert-base-uncased
RoBERTa	BooksCorpus + English Wikipedia + CC-NEWS + STORIES + OPENWEBTEXT	byte-level BPE	FacebookAI/ roberta-base
Spanish language models			
BETO	OPUS Project + Spanish Wikipedia	SentencePiece	dccuchile/ bert-base-spanish-wwm-cased
SpanBERTa	Spanish OSCAR	byte-level BPE	skimai/ spanberta-base-cased
RoBERTuito	Spanish tweets	SentencePiece	pysentimiento/ robertuito-base-uncased
Multilingual models			
mBERT	Wiki-100	WordPiece	google-bert/ bert-base-multilingual-cased
XLM-R	CC-100	SentencePiece	FacebookAI/ xlm-roberta-base

the corresponding task. Beforehand, few preprocessing steps were carried out, including converting emojis into text using the *emoji* package [5] and removing the #USER#, #URL# phrases in the original text.

The configuration (hyperparameters) we used for every model is the same for each task. Table 2 contains these values.

Table 2
The configuration for every model

Hyperparameter	Value
Optimizer	AdamW
Loss function	Cross-Entropy Loss
Pre-trained model learning rate	1e-5
classification layer learning rate	1e-4
train batch size	64
epochs	10

4. Data

This section describes the dataset provided in the challenge HOPE, part of IberLEF 2024.

In **Task 1 - Hope for Equality, Diversity, and Inclusion**, the datasets were collected and annotated from 2020 to 2023, which is an expansion of the SpanishHopeEDI dataset [26]. During the training phase of the shared task, the datasets provided included the training set and the testing set. Given a tweet, it is either labeled *HS* (Hope Speech) or *NHS* (Non-Hope Speech).

In **Task 2 - Hope as Expectations**, the tweets for English and Spanish were retrieved in the first half of 2022. They are selected through systematic filtration and annotation processes to make suitable datasets for each language [27]. In the binary classification problem, the given text is either labeled

[5] https://pypi.org/project/emoji/

Hope or *Not Hope*. In the multiclass classification problem, the class *Not Hope* is the same as binary classification while the class *Hope* is divided into *Realistic Hope*, *Generalized Hope*, and *Unrealistic Hope* to fit specific circumstances.

Table 3 displays the general statistic of the datasets, while Table 4 depicts the distribution between the classes in both tasks. As can be seen, the multiclass English dataset is imbalanced with the majority of classes focusing on *Not Hope*, while *Realistic Hope* and *Unrealistic Hope* only account for a small amount. Likewise, *Hope* class only takes about half of *Not Hope* in the binary Spanish dataset, leading to a severe imbalance between classes in the multiclass Spanish dataset.

Table 3
General statistics of each task's datasets

	Task 1		**Task 2 English**		**Task 2 Spanish**	
	Train	**Val**	**Train**	**Val**	**Train**	**Val**
Number of samples	1400	200	6192	1032	6903	1150
Number of tokens	293778	43586	1141573	184304	1009834	168763
Minimum length	8	6	9	14	7	8
Maximum length	58	58	107	83	70	62
Average length	34.85	36.88	33.00	32.07	26.32	26.26

Table 4
Distribution of classes in each task

		Not Hope	**Hope**		
			Realistic Hope	**Generalized Hope**	**Unrealistic Hope**
Task 1	Train	700	700		
	Val	100	100		
Task 2 - Binary English	Train	3088	3104		
	Val	502	530		
Task 2 - Binary Spanish	Train	4701	2202		
	Val	799	351		
Task 2 - Multiclass English	Train	3088	730	1726	648
	Val	502	300	102	128
Task 2 - Multiclass Spanish	Train	4701	505	1151	546
	Val	799	74	91	186

5. Evaluation

During the training phase of the challenge, we fine-tuned the pre-trained language models listed in Section 3 on the training datasets and then evaluated them on the validation datasets. The performance of each model was evaluated mainly by using Macro F1-score for all tasks since this is the main metric used by the organizers. This evaluation aims to assess the performance of different pre-trained language models and understand which performs best on previously unseen data. The results are shown in Table 5.

Based on the results achieved, the models with the best performance were selected to predict the classes of unseen data. Despite achieving the highest Macro F1-score, **XLM-R was only tested after the publication of the official results**. Therefore, we decided to use RoBERTuito for task 1, BERT for both English subtasks of task 2, and BETO for both Spanish subtasks of task 2.

[6]XLM-R was tested after the release of the offical results

Table 5
Evaluation of the pre-trained language models testing on validation datasets

Model	Task 1	Task 2 - Binary		Task 2 - Multiclass	
	Spanish	English	Spanish	English	Spanish
English language models					
BERT	-/-	**85.90**	-/-	**71.57**	-/-
DistilBERT	-/-	85.34	-/-	70.33	-/-
RoBERTa	-/-	85.29	-/-	69.43	-/-
Spanish language models					
BETO	81.97	-/-	**82.28**	-/-	**63.72**
SpanBERTa	85.00	-/-	80.19	-/-	58.96
RoBERTuito	**87.47**	-/-	81.70	-/-	61.21
Multilingual models					
mBERT	79.46	84.08	79.40	68.34	58.42
XLM-R[6]	83.99	**86.71**	**82.78**	**73.32**	**64.29**

6. Results

The official results and the results of the top systems are shown in Table 4. The number in brackets next to the team's name denotes their rank. Teams with equal score are the same in rank.

As it can be seen, our method helps us achieve competitive results in each task with 3rd rank or higher. Specifically, in task 1, we attained top 1 with a 0.6579 Macro F1-score, 0.0582 less than top 1's score. Regarding task 2 binary classification problem, we got 0.85 and 0.83 in Macro F1-score for English and Spanish respectively, both achieved top 3 and were 0.2 less than the F1-scores of top 1. For task 2 multiclass classification, we got top 1 for the English language with a Macro F1-score of 0.72, tied with 2 other teams, and for the Spanish language, we got top 3 with a 0.65 F1-score, lower than top 1 team's method by 0.2.

Table 6
Top 5 submissons in each task

Task 1		Task 2 - Binary English		Task 2 - Binary Spanish		Task 2 - Multiclass English		Task 2 - Multiclass Spanish	
Team	**Macro-F1**	**Team**	**Macro-F1**	**Team**	**Macro-F1**	**Team**	**Macro-F1**	**Team**	**Macro-F1**
thindang (1)	0.7161	hongson04 (1)	0.87	olp (1)	0.85	**Ours** (1)	0.72	hongson04 (1)	0.67
Ours (2)	0.6579	olp (2)	0.86	hongson04 (2)	0.84	olp (1)	0.72	olp (2)	0.66
michaelibrahim (3)	0.6522	ronghao (2)	0.86	**Ours** (3)	0.83	hongson04 (1)	0.72	**Ours** (3)	0.65
Jesus_Armenta (4)	0.6438	**Ours** (3)	0.85	ronghao (3)	0.83	ronghao (2)	0.68	MIKHAIL (4)	0.64
hongson04 (5)	0.5879	MIKHAIL (3)	0.85	KavyaG (4)	0.82	zahraahani (2)	0.67	KavyaG (4)	0.64

7. Conclusion and Future Work

In this paper, we describe our submission system for the IberLEF 2024 HOPE. After some preprocessing, we fine-tuned various pre-trained languages to fit specific tasks and concluded which models performed best on which tasks. This simple yet effective approach shows the effectiveness of pre-trained language models for downstream tasks, which helped us achieve competitive results, ranking 3rd or higher across all tasks.

However, due to resource limitations, we couldn't experiment with larger-sized models, and with myriad existing models, we couldn't try them all to validate their performance. For future work, we could explore large language models such as LLaMa [28], LaMDA [29] or Gemini [30]. Rather than simply fine-tuning language models, we could also try implementing better architecture into our models and use language models as a base for them. Furthermore, the imbalanced datasets also poses a challenge for any model to learn the pattern of all classes thoroughly. For this problem, we believe applying effective pre-processing techniques would enhance performance on the test set.

Acknowledgments

We want to show our gratitude for Dang Van Thin, our mentor for introducing us to this shared task and guiding us through many procedures.

References

[1] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, in: ECAI 2020, IOS Press, 2020, pp. 1881–1889.

[2] S. Saumya, A. K. Mishra, Iiit_dwd@ lt-edi-eacl2021: hope speech detection in youtube multilingual comments, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 107–113.

[3] D. Hardman, Pretending to care, Journal of Medical Ethics 49 (2023) 506–509.

[4] 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, 2020, pp. 41–53.

[5] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. Lambebo Tonja, J. A. García-Díaz, S. Bozkurt, B. R. Chakravarthi, H. G. Ceballos, V.-G. Rafael, G. Sidorov, L. A. Ureña-López, A. Gelbukh, S. M. Jiménez-Zafra, Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations, Procesamiento del Lenguaje Natural 73 (2024).

[6] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[7] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, et al., 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, 2022, pp. 378–388.

[8] B. R. Chakravarthi, Hope speech detection in youtube comments, Social Network Analysis and Mining 12 (2022) 75.

[9] S. Dowlagar, R. Mamidi, 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, 2021, pp. 86–91.

[10] S. Arunima, A. Ramakrishnan, A. Balaji, D. Thenmozhi, et al., ssn_dibertsity@ lt-edi-eacl2021: hope speech detection on multilingual youtube comments via transformer based approach, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 92–97.

[11] I. S. Upadhyay, A. Wadhawan, R. Mamidi, et al., Hopeful_men@ lt-edi-eacl2021: Hope speech detection using indic transliteration and transformers, arXiv preprint arXiv:2102.12082 (2021).

[12] 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).

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[14] 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).

[15] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, P. Kumar, Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4948–4961.

[16] S. M. Jiménez-Zafra, M. Á. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi,

R. Valencia-García, L. A. Ureña-López, Overview of hope at iberlef 2023: Multilingual hope speech detection, Procesamiento del Lenguaje Natural 71 (2023) 371–381.
[17] Z. Ahani, G. Sidorov, O. Kolesnikova, A. Gelbukh, Zavira at hope2023@ iberlef: Hope speech detection from text using tf-idf features and machine learning algorithms (2023).
[18] M. Shahiki-Tash, J. Armenta-Segura, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org, 2023.
[19] 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).
[20] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[21] J. Huertas-Tato, A. Martin, D. Camacho, Bertuit: Understanding spanish language in twitter through a native transformer, arXiv preprint arXiv:2204.03465 (2022).
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[23] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[24] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).
[25] 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).
[26] D. García-Baena, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgbt case, Language Resources and Evaluation 57 (2023) 1487–1514.
[27] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078. doi:10.1016/j.eswa.2023.120078.
[28] 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).
[29] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., Lamda: Language models for dialog applications, arXiv preprint arXiv:2201.08239 (2022).
[30] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).