

# Ometeotl at HOPE2024@IberLEF: Custom BERT Models for Hope Speech Detection

Jesús Armenta-Segura, Grigori Sidorov

*Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC IPN), Juan de Dios Bátiz Av., Gustavo A. Madero, 07738 Ciudad de México, México.*

## Abstract

Hope speech has the potential to mitigate hostile environments and alleviate illnesses and depression, making it crucial to detect automatically. In this paper, we present our submissions for the shared task HOPE at IberLEF 2024. This shared task encompasses two datasets: HopeEDI, consisting of tweets in Spanish, and PolyHope, which includes tweets in both English and Spanish. We proposed the use of custom BERT models, specifically pretrained on multilingual datasets and tailored for sentiment analysis. We achieved fourth, sixth and eighth place in the HopeEDI, PolyHope in Spanish, and PolyHope in English datasets, respectively, with respect of the Averaged F1 score. The code to reproduce our results can be found in <https://github.com/JesusASmx/ometeotl-Hope2024>

## Keywords

Transformers, Hope Speech Detection, Natural Language Processing, CEUR-WS

## 1. Introduction

Hope, an exceptional human trait, empowers individuals to anticipate forthcoming events and potential outcomes with adaptability. These anticipations wield significant sway over emotions, behavior, and mood [1]. Those with high levels of hope demonstrate distinctive responses to obstacles in comparison to their low-hope counterparts; they perceive impediments as challenges to surmount and utilize their cognitive pathways to devise alternative strategies toward achieving their objectives [2, 3]. Furthermore, elevated levels of hope have been linked to numerous beneficial outcomes such as academic achievement [4] and decreased levels of depression [5]. Conversely, low levels of hope are associated with adverse effects, including diminished well-being [6].

In the past two years, there has been a notable surge of interest in hope speech within the domain of Natural Language Processing (NLP). Shared initiatives focused on hope speech detection were convened at the second workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2022), a part of ACL 2022 [7], LT-EDI-2023 at RANLP 2023 [8], and the HOPE shared task at IberLEF 2023 [9]. Recently, another collaborative endeavor has emerged [10], proposed at IberLEF 2024 [11], concentrating on exploring hope from two vantage points:

---

*IberLEF 2024, September 2024, Valladolid, Spain*


✉ [jarmentas2022@cic.ipn.mx](mailto:jarmentas2022@cic.ipn.mx) (J. Armenta-Segura); [sidorov@cic.ipn.mx](mailto:sidorov@cic.ipn.mx) (G. Sidorov)

🌐 <https://jesusasmx.github.io/> (J. Armenta-Segura); <https://www.cic.ipn.mx/~sidorov> (G. Sidorov)

🆔 0009-0003-8729-7096 (J. Armenta-Segura); 0000-0003-3901-3522 (G. Sidorov)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

- Hope for equality, diversity, and inclusion [12].
- Hope as expectations [13].

The former perspective, previously investigated in the prior edition of IberLEF 2023, is further enriched in this new iteration by expanding and refining the Spanish dataset to encompass hope speech directed towards the LGTBI community. Conversely, the latter perspective remains unexplored in any prior shared task.

Detecting hope speech presents several challenges that need to be addressed to develop effective models. Firstly, there's a need to explore various machine learning approaches tailored for multilingual hope detection, considering the diverse linguistic nuances across different languages. Secondly, techniques must be devised to handle the imbalanced distribution of labels within datasets, ensuring fair representation of different hope expressions. Additionally, the noisy nature of social media data poses a challenge, requiring robust methods to filter out irrelevant information and identify genuine instances of hope speech. Moreover, extending hope detection beyond the domains in which systems are trained necessitates adapting models to recognize hope in diverse contexts. Linguistic disparities between English and Spanish texts further complicate the task, requiring specialized approaches to capture hope expressions effectively.

In order to overcome these challenges, we proposed a model based on custom BERT architectures [16] pretrained on multilingual datasets, specifically tailored for sentiment analysis. Our selection of transformers stems from their revolutionary impact on natural language processing tasks, notably through the introduction of the self-attention mechanism [17]. This mechanism empowers models to capture global dependencies and contextual relationships within sequences, making transformers highly effective in tasks such as sentiment analysis and, notably, hate speech detection [18]. Additionally, to mitigate the challenges posed by imbalanced data distribution, we adopted a strategy of selecting predictions based on higher logits relative to the less represented class. This approach significantly improved results, underscoring the importance of the challenges highlighted by the organizers of the shared task. For instance, while submissions in the HopeEDI dataset [12] using argmax predictions achieved a best average F1-score of 0.51, our latest submission achieved a score of 0.64, only three points below the top-ranked submission. The code to reproduce these results can be found in <https://github.com/JesusASmx/ometeotl-Hope2024>

The structure of this paper is as follows: in Section 2, we detail our methodology. In Section 3, we outline our experimental workflow. In Section 4, we discuss the results of our experiments. Finally, in Section 5, we conclude the paper.

## 2. Methodology

Our main aim was to create a technique that could embed the hope short texts into a two-dimensional framework, delineating between *Hope* and *Not Hope*. This was accomplished by utilizing a loss function and a backpropagation algorithm to adjust these values towards

the provided golden labels. As explained in the introduction, we employed custom BERT architectures pretrained on multilingual datasets tailored for sentiment analysis. Specifically, we leveraged the Huggingface model *nlptown/bert-base-multilingual-uncased-sentiment* [19].

We also employed k-fold cross-validation in order to ensure the robustness of our predictions. We stated  $k = 5$  and we splitted the training set into 5 parts. For each epoch, we trained the model  $k$  times by taking as validation set each one of the  $k$  parts. In all cases, the learning curves were reasonable fitted by considering the high complexity of the datasets. As a consequence, this election was correct.

### 3. Experimental Setup

#### 3.1. Data

For the three datasets [12, 13], two were in Spanish and one in English. The HopeEDI dataset contains 2, 550 tweets, while the PolyHope in Spanish contains 28, 424 tweets. The PolyHope dataset in English contains 28, 424 tweets. See Table 3.1 for the full statistics of each datasets.

Language	Split	Hope Speech	Not Hope Speech	Total
HopeEDI (Spanish)	Train data	700	700	1, 400
	Validation data	100	100	200
	Test data	200	200	400
PolyHope (Spanish)	Train data	2, 202	4, 701	6, 903
	Validation data	351	799	1, 150
	Test data	379	773	1, 152
PolyHope (English)	Train data	3, 088	3, 104	6, 192
	Validation data	502	530	1, 032
	Test data	491	541	1, 032

**Table 1**

Statistics of the datasets. The HopeEDI dataset was balanced, while both PolyHope datasets were not.

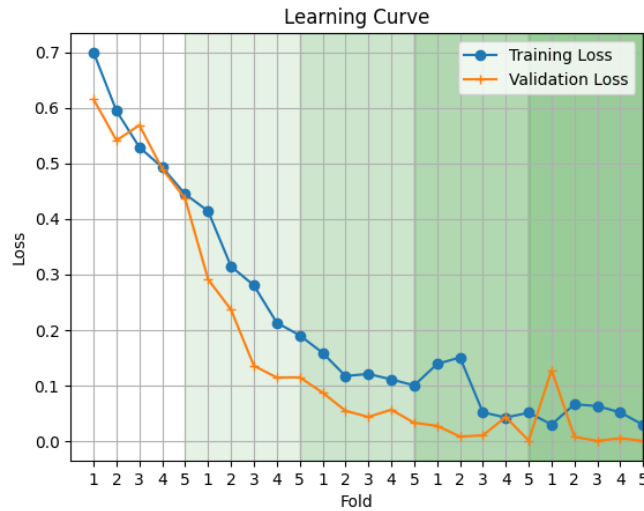
#### 3.2. Experimental workflow

In the first phase of the contest, we experimented with several custom BERT models over the validation set. Since the model *nlptown/bert-base-multilingual-uncased-sentiment* presented the best results, we used it for the final phase of the contest. For all experiments, the employed hiperparameters were:

- **Random Seed:** 42.
- **Folds for Cross Validation:** 5.
- **Epochs:** 5.
- **Batch Size:** 32.
- **Max. tokenizer length:** 128.

- **Token Padding system:** *Left*.
- **Padding Token:** *Eos Token*.
- **Label encoding:**
  - **PolyHope:** *Not Hope = 0* and *Hope = 1*.
  - **HopEDI:** *NHS = 0* and *HS = 1*.
- **Optimizer:** Adam.
- **Learning Rate:**  $5e - 5$ .
- **Epsilon value:**  $1e - 8$ .

Figures 1 and 2 depicts the evolution of the performance through the epochs, in the three experiments.

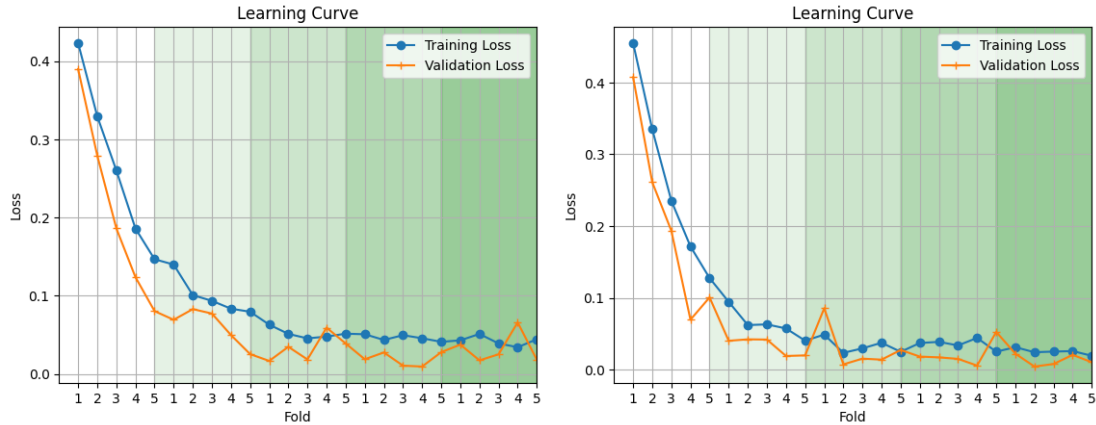


**Figure 1:** Evolution of the experiments over the HopeEDI dataset through the 5 epochs. For each epoch,  $k = 5$  and it is determined with the gradient of the green. Underfitting is presented in epoch 1, fold 3 and in epoch 5, fold 2. The rest of the training had a reasonable fitting by considering the monotonicity and the complexity of the data.

## 4. Results and Discussion

In terms of the average F1 score, we obtained 0.64 in the HopeEDI dataset, only 0.03 points under the first place. We obtained 0.84 in the PolyHope Spanish dataset, only 0.03 points under the first place. Finally, we obtained 0.82 in the PolyHope Spanish dataset, 0.05 points under the first place. The full leaderboards can be consulted in Table 2 for the HopeEDI dataset, Table 3 for the PolyHope Spanish dataset and Table 4 for the PolyHope English dataset.

All the issues warned by the organizers revealed to have a deep impact in any trained model, and our decisions to rely in multilingual models with cross-validation training demonstrated to



**Figure 2:** Evolution of the experiments over the PolyHope (EN, left) and PolyHope (ES, right) datasets through the 5 epochs. For each epoch,  $k = 5$  and it is determined with the gradient of the green. For the English dataset, underfitting is presented in epoch 3, fold 4 and in epoch 5, fold 4. The rest of the training had a reasonable fitting by considering the monotonicity and the complexity of the data. For the Spanish dataset, underfitting is presented in epoch 3, fold 2 and in epoch 4, fold 5. The rest of the training had a reasonable fitting by considering the monotonicity and the complexity of the data.

be a strong proposition, by considering how close we were from the first place. Moreover, our custom prediction algorithm explained in the introduction to perform the predictions, allowed us to overcome the problem of data unbalance.

Pos.	Team	Macro F1	Weighted Macro F1
1	thindang	0.67	0.67
2	ChauPhamQuocHung	0.66	0.66
3	hongson04	0.65	0.65
4	<b>Us</b>	<b>0.64</b>	<b>0.64</b>
5	michaelibrahim	0.64	0.63
	.....		
15	Arunraj_Subburaj	0.54	0.54
16	MIKHAIL	0.51	0.51

**Table 2**

Evaluation results for the HopeEDI dataset. We achieved fourth place out of 16 participants.

With respect of the learning curves depicted in Figures 1 and 2, both cases shows an overall good fit along with a adequate decreasing rate. However, there are instances indicating issues with model performance. In the case of the HopeEDI dataset, underfitting is observed in epoch 1, fold 3, where the model fails to capture the underlying patterns in the training data, but overfitting is noted in epoch 5, fold 2, where the model may lacks of generalization capabilities with respect of the training data.

In the case of the PolyHope datasets, the english case presents underfitting in epoch 3, fold 4, and in epoch 5, fold 4. In these instances, the model fails to adequately learn from the training

Pos.	Team	Macro F1	Weighted Macro F1
1	olp	0.85	0.87
2	hongson04	0.84	0.86
3	ChauPhamQuocHung	0.83	0.85
4	ronghao	0.83	0.85
5	KavyaG	0.82	0.84
6	<b>Us</b>	<b>0.81</b>	<b>0.84</b>
7	MIKHAIL	0.81	0.83
... ..			
13	thindang	0.73	0.76
14	Fida	0.71	0.74

**Table 3**

Evaluation results for the PolyHope Spanish dataset. We achieved sixth place out of 14 participants.

Pos.	Team	Macro F1	Weighted Macro F1
1	hongson04	0.87	0.87
2	olp	0.86	0.86
3	ronghao	0.86	0.86
4	ChauPhamQuocHung	0.85	0.85
5	MIKHAIL	0.85	0.85
6	zahraahani	0.85	0.85
7	hamadanayel	0.83	0.83
8	<b>Us</b>	<b>0.82</b>	<b>0.82</b>
9	KavyaG	0.82	0.81
... ..			
16	AmnaNaseeb	0.23	0.31
17	JuanCalderon	0.21	0.21

**Table 4**

Evaluation results for the PolyHope English dataset. We achieved eighth place out of 17 participants.

data, resulting in high training and validation errors, indicating that the model is too simplistic and does not capture the underlying patterns in the data. For the spanish case, underfitting is evident in epoch 3, fold 2, and in epoch 4, fold 5.

## 5. Conclusions and Further Work

Detecting hope speech presents several challenges that must be addressed to develop effective models capable of operating across diverse linguistic landscapes and social media multilingual contexts. Additionally, handling imbalanced label distributions within datasets is essential to ensure fair representation of diverse hope expressions. Furthermore, the noisy nature of social media data necessitates the development of robust methods to filter out irrelevant information and identify genuine instances of hope speech amidst the clutter.

In this paper, custom BERT models were proposed to address these challenges, pretrained on

multilingual datasets for sentiment analysis. We performed the training with 5-cross-validation and we adopted a strategy of selecting predictions based on higher logits relative to the less represented class to mitigate imbalanced data distribution issues.

Our experimental results, very close to the first place of each category, demonstrate the effectiveness of our proposed approach. Concretely, we achieved weighted macro F1-scores of 0.64 in the HopeEDI dataset (only 0.03 points behind the top place); 0.84 in the PolyHope Spanish dataset (only 0.05 points behind the top place) and 0.82 in the PolyHope Spanish dataset (only 0.05 points behind the top place).

Further work may consist in exploring more robust multilingual custom BERT models. For instance, this same *nlptown/bert-base-multilingual-uncased-sentiment* model can be pretrained again with more datasets in more languages in order to enhance its ability to generalize across different linguistic contexts. Concretely, it can be expanded with other Spanish or Latin-language datasets. Another aspect with room for enhancement is the manipulation of the datasets itself. For instance, it may be possible to explore data augmentation in order to evaluate its viability to enhance the results. Conversely, data reduction aiming the elimination of noise can be useful: Developing more advanced filtering algorithms to clean social media data will be crucial. Employing noise reduction strategies such as anomaly detection or clustering can help in isolating genuine instances of hope speech from irrelevant or misleading content.

By addressing these areas, future research can build on the foundation laid by this study to develop more robust, accurate, and generalizable models for detecting hope speech across diverse and noisy social media landscapes.

## References

- [1] P. Bruininks, B. F. Malle, Distinguishing hope from optimism and related affective states, *Motivation and Emotion* 29 (2005) 324–352.
- [2] C. R. Snyder, *The Psychology of Hope: You Can Get There from Here*, Simon and Schuster, 1994.
- [3] C. R. Snyder, Hypothesis: There is hope, in: *Handbook of Hope*, Elsevier, 2000, pp. 3–21.
- [4] C. R. Snyder, Hope theory: Rainbows in the mind, *Psychological Inquiry* 13 (2002) 249–275.
- [5] C. R. Snyder, B. Hoza, W. E. Pelham, M. Rapoff, L. Ware, M. Danovsky, L. Highberger, H. Ribinstein, K. J. Stahl, The development and validation of the children’s hope scale, *Journal of Pediatric Psychology* 22 (1997) 399–421.
- [6] E. Diener, Subjective well-being, in: *The Science of Well-Being*, 2009, pp. 11–58.
- [7] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. C. Navaneethakrishnan, J. P. McCrae, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, R. Valencia-García, P. K. 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: *Association for Computational Linguistics*, 2022, pp. 378–388.
- [8] P. Kumaresan, B. R. Chakravarthi, S. Cn, M. Á. García, S. M. Jiménez-Zafra, J. A. García-Díaz,

- R. Valencia-García, M. Hardalov, I. Koychev, P. Nakov, D. García-Baena, K. K. Ponnusamy, B. Preston, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2023, pp. 47–53.
- [9] S. M. Jiménez-Zafra, M. Á. García-Cumbreras, D. García-Baena, J. A. García-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.
- [10] 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).
- [11] 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.
- [12] 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.
- [13] 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.
- [14] J. Armenta-Segura, G. Sidorov, Anime Success Prediction Based on Synopsis Using Traditional Classifiers, in: Proceedings of Congreso Mexicano de Inteligencia Artificial, COMIA, 2023.
- [15] J. Armenta-Segura, G. Sidorov, Anime popularity prediction before huge investments: a multimodal approach using deep learning, 2024. URL: <https://arxiv.org/abs/2406.16961>. arXiv:2406.16961.
- [16] 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).
- [17] 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).
- [18] J. Armenta-Segura, C. J. Núñez-Prado, G. O. Sidorov, A. Gelbukh, R. F. Román-Godínez, Ometeotl@multimodal hate speech event detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained BERT models over text, in: A. Hürriyetoğlu, H. Tanev, V. Zavarella, R. Yeniterzi, E. Yörük, M. Slavcheva (Eds.), Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 53–59. URL: <https://aclanthology.org/2023.case-1.7>.
- [19] Nlptown bert model for multilingual sentiment analysis, Available at: .



<https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>.