UMUTeam at HOPE2023@IberLEF: Evaluation of Transformer Model with Data Augmentation for Multilingual Hope Speech Detection

Ronghao Pan[1,*,†], Gema Alcaraz-Mármol[2,†] and Francisco García-Sánchez[1,†]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Departamento de Filología Moderna, Universidad de Castilla La Mancha, 45071, Spain

Abstract

This paper describes the participation of the UMUTeam in the HOPE shared task organized at IberLEF 2023 within the SEPLN conference. We have addressed the two proposed subtasks. The objective of both subtasks is the detection of hopeful speech texts, but in the first subtask the texts are in Spanish and in the second one in English. The approach presented for both subtasks is based on fine-tuning different pre-trained Large Language Models (LLMs) based on Transformer with data augmentation for the sequence classification task, particularly for hope speech detection. In subtask 1, our team ranked in fifth position out of 11 participants, with a macro f1 score of 71.03, while in subtask 2 we were placed in seventh position out of 9 participants, with a macro f1 score of 48.22.

Keywords

Natural Language Processing, Transformers, Hope speech detection, Large Language Model

1. Introduction

Hope Speech is a type of speech that is able to relax a hostile environment and is designed to inspire and motivate people to take positive action towards achieving a particular goal or overcoming a challenge [1]. One of the primary goals of hope speech is to give suggestions inspiring for good to a number of people when they are in adversity or challenging circumstances [2]. Therefore, automatic detection of hope speech and positive comments in the text can be a powerful tool to combat sexual or racial discrimination and foster positive environments [1]. People are targeted with offensive messages on social media due to their race, color, ethnicity, gender, sexual orientation, nationality, or religion. According to [2], the significance of social media in the lives of vulnerable groups like the LGBT community, racial minorities, and people with disabilities has been studied, revealing that an individual's social media activities can significantly shape their personality and worldview. The issue of offensive messages on social media is a widely discussed topic across various languages and platforms [3].

The aim of the *HOPE: Multilingual Hope Speech detection* shared-task [4], as part of the IberLEF 2023 [5] workshop within the framework of the 39th International Conference of the Spanish

IberLEF 2023, September 2023, Jaén, Spain

ronghao.pan@um.es (R. Pan); gema.alcaraz@uclm.es (G. Alcaraz-Mármol); frgarcia@um.es (F. García-Sánchez)
0009-0008-7317-7145 (R. Pan); 0000-0001-7703-3829 (G. Alcaraz-Mármol); 0000-0003-2667-5359 (F. García-Sánchez)

Society for Natural Language Processing (SEPLN 2023) is the detection of the discourse of hope, in pursuit of equality, diversity, and inclusion. The organizers proposed two subtasks. The first one is a binary classification to identify whether a text in Spanish contains hope speech or not. The second one has the same objective as the first one, but in this case, the texts are YouTube comments in English.

This work presents the participation of UMUTeam in both subtasks, which is based on the fine-tuning of different pre-trained Large Language Models (LLMs) based on Transformer [6] with data augmentation. The rest of the paper is organized as follows. Section 2 presents the task and dataset provided. Section 3 describes the methodology of our proposed system for addressing subtask 1 and subtask 2. Section 4 shows the results obtained. Finally, Section 5 concludes the paper with some findings and possible future work.

2. Task description

The HOPE shared task, organized at IberLEF 2023 workshop, aims to detect and identity hope speech in two languages, Spanish and English, and two different social networks, namely, Twitter and YouTube. Specifically, the organizers propose two tasks for this challenge:

- **Subtask 1**: Determine if a given Spanish tweet contains hope speech or not.
- **Subtask 2**: Detect whether an English YouTube comment contains hope speech or not.

This shared task was previously organized at the 2nd Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2022), as part of ACL 2022, but for five languages: Tamil, Malayalam, Kannada, English and Spanish. In this version, an improved and expanded dataset has been provided in English [2] and Spanish [7] and is directed to the IberLEF community. The statistics of the dataset, grouped by each subtask, are shown in Table 1. It can be observed that the dataset for subtask 1 (Spanish) is balanced, and the English dataset is not. In case of subtask 2, the proportion in train dataset between hope speech (*HS*) and non hope speech (*NHS*) is near to 1:11. In order to partially mitigate this imbalance problem, the dataset has been extended by using the hopeful speech texts from the Spanish dataset. Thus, the augmented training dataset for subtask 2 has 2 574 hopeful texts and 18 577 non hope speech texts, which constitutes a 1:7 ratio. In addition, the data augmentation technique has been employed on the Spanish dataset (subtask 1) using the hope speech texts from the English dataset to test whether it improves the performance of the model in detecting hope speech for such subtask.

It is worth mentioning that the organizers only supplied the training and testing sets, and as a result, we had to develop our own validation split. To ensure a balanced labeling, we used stratified sampling to create a custom validation split. Lastly, it's noteworthy that the organizers utilized Precision, Recall, and F1 scores to evaluate the participants' systems. These scores will be calculated for each category and averaged using the macro-average method. The macro-F1 score will be used to rank the systems.

Table 1
Distribution of the datasets.

Dataset	Total	HS	NHS
Spanish			
Train	1,289	633	656
Train (Augmented)	3,518	2,862	656
Validation	323	158	165
Test	450	150	300
English			
Train	20,360	1,783	18,577
Train (Augmented)	21,151	2,574	18,577
Validation	5,090	446	4,644
Test	4,805	21	4,784

3. Methodology

For solving this shared task, we built the system whose architecture is depicted in Figure 1. In a nutshell, our system works as follows. First, the dataset is pre-processed as described in Section 3.1. Then, the dataset is divided into training, evaluation, and testing using the strategy presented in Section 3.2. Finally, the fine-tuning of different multilingual and monolingual pre-trained models for the classification of hope speech was carried out as shown in Section 3.3.

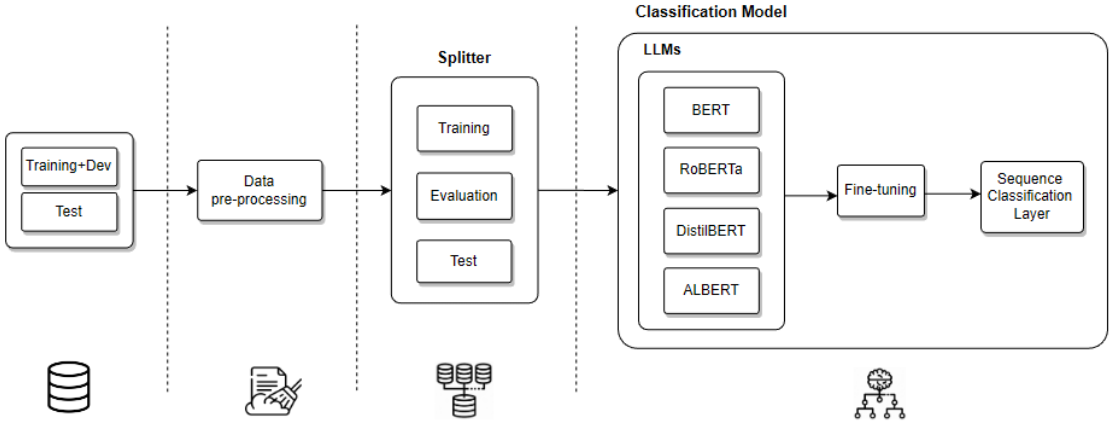

Figure 1: Overall system architecture.

3.1. Dataset preprocessing

Our preprocessing stage consists of the following processes to clean the text items in the Hope Speech dataset:

- Replacement of all hashtags and mentions with *#[HASHTAG]* and *@[USER]*.
- Social media posts require a lot of cleaning up, but it is inefficient to clean up each post, so a general clean-up approach was applied in this case:
  - All emojis have been replaced by their textual meaning, using the *emoji* library.
  - For English, general contractions have been expanded through the *contraction* library, such as "*lmao*" to "*laughing my ass off*", "*y'all*" to "*you all*", "*i'd*" to "*i would*", etc.

3.2. Splitter

As mentioned in Section 2, we employed the data augmentation technique to enhance the overall performance of our classification models. To this end, we translated the texts classified as hope speech in the train and validation splits of the English dataset into Spanish for subtask 1, and the other way around for subtask 2 (i.e., the texts classified as hope speech in the train and validation splits of the Spanish dataset were translated into English). It should be noted that the organizers provided only the training and test sets, so we had to create our own validation split. To achieve a balanced labeling, we used stratified sampling to create a customized validation split.

3.3. Classification model

We utilized the fine-tuning technique with pre-trained models such as BERT, RoBERTa, ALBERT, and DistilBERT to develop our classification model for both subtask 1 and subtask 2. Fine-tuning involves adapting a pre-trained model to a labeled dataset, allowing us to benefit from high-quality language representations that the pre-trained model has already learned for the classification task. In [8], the performance of this approach on the task of sentiment classification in Spanish financial texts has been demonstrated. For the classification of hope speech texts, we incorporated an additional layer, called *Sequence Classification Layer*, at the end of the LLMs, which allows the system to output classification results.

The pre-trained models used for subtasks 1 and 2 are as follows:

- **BETO**: It is a BERT-based model trained exclusively on a large corpus of Spanish. BETO was trained with the Whole Word Masking technique. It has been shown that by fine-tuning this model better results are obtained than with other BERT-based models pre-trained on multilingual corpora for most of the tasks, even achieving a new state-of-the-art on some of them [9].
- **ALBETO**: LLMs have made significant strides in recent years. Many approaches focus on increasing model size by pre-training natural language representation to enhance performance on downstream tasks. However, as the model size increases, GPU/TPU memory limitations and longer training times make it increasingly challenging [10]. This results in impractical inference times for real-world applications. To tackle this issue, ALBERT [10] proposes two parameter-reduction techniques to reduce memory consumption and speed up training for BERT. In this study, we utilized ALBETO, which is a Spanish corpus exclusive pre-trained version of ALBERT [11].

- **DistilBETO**: It is a DistilBERT-based model trained on a Spanish corpus. This model uses the distillation technique to transfer the knowledge of the BETO model to this new model [11].
- **MarIA**: It is based on the RoBERTa base model and has been pre-trained using the largest Spanish corpus known to date, with a total of 570GB of clean and deduplicated text, compiled from the web crawlings performed by the National Library of Spain from 2009 to 2019 [12].
- **XLM-R**: It is a multilingual version of RoBERTa pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages [13].
- **BERT**: It is a transformers model pre-trained on a large corpus of English data using a Masked Language Modeling (MLM) objective [14].
- **RoBERTa-large**: This is a large version of RoBERTa and is a transformers model pre-trained on a large corpus of English data in a self-supervised fashion [15].
- **ALBERT**: It is a transformers model pre-trained on a large corpus of English data in a self-supervised fashion and proposes two parameter-reduction techniques to reduce memory consumption and speed up training for BERT [10].
- **DistilBERT**: This model is a distilled version of the BERT base model and is smaller and faster than BERT, which was pre-trained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher [16].
- **DeBERTa-large**: This model is an improved version of BERT and RoBERTa, which uses disentangled attention and an improved mask decoder [17].

4. Results

This section describes the systems submitted by our team in each run and the overall results obtained in subtask 1 and subtask 2. It should be noted that each participating team was allowed to submit 10 runs. All models are fine-tuned with a training batch size of 16, 15 epochs, a learning rate of 2e-5, and a weight decay of 0.01.

4.1. Subtask 1

The results of each pre-trained model for subtask 1 are shown in Table 2. As can be observed, lightweight models such as ALBETO and DistilBETO have obtained the worst results with a macro f1-score of 62.07% and 57.46%, respectively. The best result was obtained with the MarIA model with a macro f1-score of 68.50%. It is also observed that monolingual models perform better than multilingual models, such as XLM-R.

To improve the overall performance of the model, we fine-tuned the best performing model (MarIA) with the augmented dataset. Table 3 shows the results obtained. We can see that the fine-tuned MarIA model with the augmented training set has improved mainly in hope speech detection and has improved overall by 2.521% in macro f1-score.

The official leaderboard for subtask 1 is depicted in Table 4. We achieved the fifth position in the ranking with a macro f1-score of 71.03%. We can also observe that our approach has a high accuracy in detecting hope speech with an F1 HS score of 64.14%, which is the fourth best result.

Table 2
Individual results of each pre-trained model without data augmentation for subtask 1. For each model, the macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1) are reported.

Model	M-P	M-R	M-F1
BETO	0.80185	0.67000	0.68324
ALBETO	0.77375	0.62167	0.62071
DistilBETO	0.82335	0.59333	0.57469
MarIA	0.68371	0.68667	**0.68508**
XLM-R	0.78172	0.65667	0.66695

Table 3
Result of MarIA with data augmentation.

	Precision	Recall	F1-score
HS	0.56995	0.73333	0.64140
NHS	0.84436	0.72333	0.77917
Macro avg	0.70715	0.72833	**0.71029**

The teams *haanh764* and *JL_DomOlmedo* outperformed our best run with a macro F1-score (M-F1) of 91.61% and 74.37%, respectively. As commented above, we achieved this result using the fine-tuned MarIA model with data augmentation.

Table 4
Official leaderboard for subtask 1

#	Team Name	M-F1	F1 HS	F1 NHS
1	haanh764	0.9161	0.8896	0.9426
2	JL_DomOlmedo	0.7437	0.6167	0.8707
3	zahraahani	0.7430	0.6728	0.8133
4	moeintash	0.7238	0.6569	0.7907
5	**UMUTeam**	**0.7103**	**0.6414**	**0.7792**
6	honghanhh	0.7034	0.5617	0.8451
-	-	-	-	-
11	mgraffg	0.4198	0.0588	0.7808

4.2. Subtask 2

The result of subtask 2 with different pre-trained models is shown in Table 5. As can be seen, the best result has been obtained with the DeBERTa model, achieving a macro f1-score of 48.224%. In this case, one of the lightweight models, namely DistilBERT, has obtained better results than other more complex models such as BERT and RoBERTa.

In this case, due to our GPU limitations, it was not possible to fine-tune DeBERTa-v2 with the augmented dataset. Therefore, we fine-tuned BERT and RoBERTa (in post-evaluation) with

Table 5
Individual results of each pre-trained model without data augmentation for subtask 2. For each model, the macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1) are reported.

Model	M-P	M-R	M-F1
BERT	0.50295	0.56533	0.48004
RoBERTa-large	0.50520	0.62933	0.47979
ALBERT	0.50203	0.55341	0.47169
DistilBERT	0.50117	0.52346	0.48008
DeBERTa-large	0.50407	0.58955	**0.48224**

the augmented training set. The results are shown in Table 6 and it can be seen that the model has improved by 1,314% over the best model (DeBERTa).

Table 6
Individual results of each pre-trained models with data augmentation for subtask 2. For each model, the macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1) are reported.

Model	M-P	M-R	M-F1
BERT	0.50417	0.65823	0.45239
RoBERTa-large (post-evaluation)	0.50366	0.54394	**0.49538**

Table 7 depicts the official leaderboard for subtask 2. A total of nine participants sent their results. We achieved the seventh position in the official leaderboard with our run based on fine-tuning DeBERTa-large model. In addition, our approach has performed well in detecting hope speech texts with a 2.23% in F1 HS, outperforming second-place teams. In the post-evaluation phase, we have evaluated the performance of the RoBERTa-large fine-tuned with the augmented training dataset (see Table 6), and it can be observed that the macro f1-score obtained would rank fifth in the official leaderboard.

Table 7
Official leaderboard for subtask 2

#	Team Name	M-F1	F1 HS	F1 NHS
1	JL_DomOlmedo	0.5012	0.0301	0.9724
2	juanmanuel.calvo	0.4989	0.0000	0.9978
3	zahraahani	0.4975	0.0000	0.9950
4	moeintash	0.4974	0.0000	0.9949
5	varsha2010399	0.4937	0.0000	0.9875
6	honghanhh	0.4862	0.0246	0.9478
7	**UMUTeam**	**0.4822**	**0.0223**	**0.9421**
8	mgraffg	0.4651	0.0292	0.9009
9	haanh764	0.4429	0.0128	0.8730

5. Conclusion

These working notes summarize the participation of the UMUTeam in the HOPE shared task (IberLEF 2023). We participated in the two challenges proposed, achieving promising results in both. Specifically, we ranked the 5/11 in subtask 1, a binary classification task for detecting hope speech in Spanish, with a macro f1-score of 71.03%, and 7/9 in the subtask 2, a binary classification task for detecting hope speech in English, with a macro f1-score of 48.22%. For both subtasks, we used the same approach, which is based on fine-tuning different pre-trained models with data augmentation for the sequence classification task, specifically for hope speech detection.

As future work, we are planning to improve our pipeline using an expanded LLMs model with hope related speech texts, i.e. extend a Masked Language Model (MLM) model with hope speech text and later fine-tune this model for detecting hope speech texts. In addition, we are planning to incorporate the features related to figurative language into our pipeline [18]. We consider that these kinds of features can improve the overall of the system in which the words in a text differ from their literal meaning.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

[1] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, in: European Conference on Artificial Intelligence, 2019.

[2] 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.

[3] M. Mozafari, R. Farahbakhsh, N. Crespi, Cross-lingual few-shot hate speech and offensive language detection using meta learning, IEEE Access 10 (2022) 14880–14896. URL: https://doi.org/10.1109/ACCESS.2022.3147588. doi:10.1109/ACCESS.2022.3147588.

[4] 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).

[5] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, 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.

[6] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMUS : A survey of transformer-based pretrained models in natural language processing, CoRR abs/2108.05542 (2021). URL: https://arxiv.org/abs/2108.05542. arXiv:2108.05542.

[7] 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 (2023) 1–28.

[8] J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Smart analysis of economics sentiment in spanish based on linguistic features and transformers, IEEE Access 11 (2023) 14211–14224. URL: https://doi.org/10.1109/ACCESS.2023.3244065. doi:10.1109/ACCESS.2023.3244065.

[9] 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.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[11] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and distilbeto: Lightweight spanish language models, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 4291–4298. URL: https://aclanthology.org/2022.lrec-1.457.

[12] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[13] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[15] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[16] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[17] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[18] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex &

Intelligent Systems 8 (2022) 1723–1736.