URJC-Team at HOPE2023@IberLEF: Multilingual Hope Speech Detection Using Transformers Architecture

Miguel Ángel Rodríguez-García[1,*], Adrián Riaño-Martínez[1] and Soto Montalvo Herranz[1]

[1]*Universidad Rey Juan Carlos, Spain*

Abstract

Detecting Hope Speech refers to identifying content in natural language that provokes optimism in people's minds, encouraging them to improve their life. This type of speech has a relevant target in our society, offering supporting messages for people suffering from depression, stress and loneliness. Despite its relevance, a significant number of published studies address to recognise the flip side of the coin, hate speech. This work describes our contribution to the HOPE challenge, i.e., detecting hope speech content in Spanish and English texts. Two different transform models are proposed to tackle the subtasks suggested in the challenge. Our proposal reaches notable results.

Keywords

Deep Learning, Transformers, Natural Language Processing, Hope Speech

1. Introduction

According to health experts hope is essential for human health [1]. It is an intrinsic part of human life and vital for enhancing the quality of life. Hope is conceived as a motivational resource capable of increasing happiness and decreasing bad feelings such as stress and helplessness [2, 3]. Its detection in natural language has a direct impact on product reviews, election campaigns, or decision-making in various contexts [4]. It can be employed to diffuse hostility in social media [5], reduce enmity in politics during a conflict [6], and inspire people suffering from depression, loneliness and stress [7]. Besides, its application on social media platforms has supposed a direct pathway for linguistics computer scientists and psychologists to dive deep into multiple ways of human interactions [8].

Due to its relevancy, there is a need for automatically detecting hope-speech in natural language. In this sense, Artificial Intelligence influences this classification problem through language technologies. Deep and Machine Learning stand for the most widely used technologies to build statistical models capable of interpreting the sequence of words and accurately solving

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

miguel.rodriguez@urjc.es (M. Á. Rodríguez-García); a.riano.2016@alumnos.urjc.es (A. Riaño-Martínez); soto.montalvo@urjc.es (S. M. Herranz)

0000-0001-6244-6532 (M. Á. Rodríguez-García); 0009-0004-8755-255X (A. Riaño-Martínez); 0000-0001-8158-7939 (S. M. Herranz)

specific problems in Natural Language Processing, such as machine translation [9], speech recognition [10], and text summarization [11], among others.

In this paper, we describe the approach submitted to the IberLEF 2023 task HOPE - Multilingual Hope Speech detection [12, 13]. We propose a system based on Deep Learning architectures. Thus, we have experimented with cutting-edge architectures like Encoders-Decoders and their families of masked-language models. Concretely, the proposed approach for the challenge is fundamentally based on Transformers, Bidirectional Encoder Representations from Transformers (BERT) for English tasks and its version trained on Spanish corpus, BETO. Both choices were selected as the primary tool to address the challenge due to, after the literature review, we observed that these kinds of models were the ones which reached the highest results.

The rest of the paper is organised as follows. Section 2 presents an overview of the related work. Section 3 details the distribution of the datasets delivered for each task and describes the Deep Learning architecture proposed for each one. Section 4 presents the results achieved in the challenge. Finally, Section 5 depicts the main conclusions of work.

2. Related Work

Studying the speech has largely attracted the researchers' attention since it reveals a range of various information related to, for instance, the emotional state [14]. According to the literature, the speech has been widely analysed in numerous ways for identifying the negative effects, determining abusive language that incites violence, or the positive one, identifying encouraging comments that provokes reassurance. In this work, we focused on hope speech, due to the relative attention acquired.

In [15], Chakravarthi created a multilingual dataset compiled from comments on Youtube in different languages and experimented with two kinds of statistical models to identify positivity. One of this type is a variety of Machine Learning models, and the other consisted of a deep neural network architecture. As a result, the author concluded that the proposed CNN model outperforms others. Following a similar method, Saumya and Mishra in [16] describe the system submitted to a similar challenge [17]. They employed the same social media source, Youtube, and various Machine Learning, Deep Learning and hybrid models to identify Hope Speech in English, Tamil, and Malayalam languages. The results revealed that in the majority of the conducted experiments, Deep Learning models slightly outperform the conventional Machine Learning models. However, the best results were achieved by hybrid learning models defined by Long short-term memory (LSTM) architecture and its bidirectional version BiLSTM. Another interesting submission was Ghanghor et al., [18], where they proposed a solution based on customized versions of several transformed-based pre-trained models. Frozing the models, modifying the loss functions, and redefining the last layers were some of the changes conducted by the authors. Experiments with the test datasets revealed that pre-trained models without customization reached better performance than other approaches, above all in the English language. Out of this challenge, Balouchzahi et al., in [19], reported an English hope speech dataset compiled from Twitter. The dataset was benchmarked using three different baselines based on traditional Machine Learning, Deep Learning and Transformers. In spite of the past two belonging to the same family, Deep Learning, the authors differ on both baselines

considering the employed architecture, CNN and BiLSTM architecture as a Deep Learning and Encoder-Decoder structure as a Transformers. Besides, the presented approach detail the strict annotation process followed to classify the dataset, and the experiments carried out benchmarking it. The results showed that transformer models reached better performance compared to other approaches. In conclusion, after the analysis conducted, due to the notable results obtained by transformers across the literature review study, we decided to take this strategy to overcome the tasks delivered in the challenge.

3. Material and Methods

This section details the dataset delivered in the challenge and the methods employed to face the tasks submitted.

3.1. Data

The dataset provided by the organizers is formed by two corpora, Spanish and English compiled from different social media platforms. The former was compiled from Twitter and consisted of LGBT-related tweets annotated by a set of determined rules to classify them, such as positive talks about the LGTBI community in determined circumstances [20]. The latter was harvested from Youtube and consisted of comments posted on Youtube about varous relevant social topics like Diversity, Equally and Inclusion [5, 21]. Both corpora are labelled by using two tags HS (Hope Speech) and NHS (Non-Hope Speech). Following the same design, the organizers provided for each phase, training and testing phase, a different dataset. Table 1 shows the distribution of both datasets considering the number of samples for each type label.

Table 1
Development and Evaluation datasets distribution.

		Development		Evaluation	
	Label	Train	Test	Train	Test
Spanish	HS (Hope Speech)	691	300	791	150
	NHS (Non Hope Speech)	621		821	300
	Total	*1.312*	*300*	*1.612*	*450*
English	HS (Hope Speech)	1.961	2.799	2.229	21
	NHS (Non Hope Speech)	20.690		23.221	4.784
	Total	*22.651*	*2.799*	*25.450*	*4805*

The rows provide the number of tweets collected on each dataset for each phase. Thus, in the Development phase, the Spanish dataset comprises 1.312 tweets, from which 691 are classified as a Hope Speech (HS) and 621 as Non-Hope Speech (NHS). Similarly, the English corpus contains 22.651 comments, where 1.961 were labelled as HS and 20.690 as NHS. Note in the Test only appears a number since the datasets were provided without labels. Regarding

the Evaluation phase, in the Spanish task, the train and test datasets compiled 1.612 and 450 tweets, respectively. Likewise, in the English, were 25.450 and 4.805. Analysing the datasets' distribution, it is worth stressing the unbalanced amount of samples in the English dataset with respect to the Spanish.

3.2. Method

The HOPE challenge of the IberLEF evaluation campaign consisted of two different tasks about detecting hope in two languages, English and Spanish. We experimented with the relatively new sphere of Deep Learning, which is being applied to practice in a high percentage of natural language classification problems, the transformer architecture, and more precisely, one of its influential transformer models trained as a language model, the Bidirectional Encoder Representations from Transformers (BERT). In particular, BERT is a Transformer model that has been pre-trained primarily for two tasks: language modelling, where the target is to predict tokens given their context, and next sentence prediction, focused on text sequence classification tasks [22]. Similarly, in the Spanish task, we employed the Spanish version of BERT, BETO, a model based on BERT-Base architecture, a small-size model implemented, which was pre-trained with a corpus of 3 billion words created from Spanish texts extracted from Wikipedia and OPUS Project [23]. Both models were configured similarly using a batch_size, number of epochs and a learning rate of 64, 5, and 3e-5, respectively. The two models were integrated into a pipeline to address the tasks demanded in the challenge. This pipeline consisted of four steps, namely, preprocessing, building the selected model, training and testing. Figure 1 depicts a graphic representation of this pipeline.

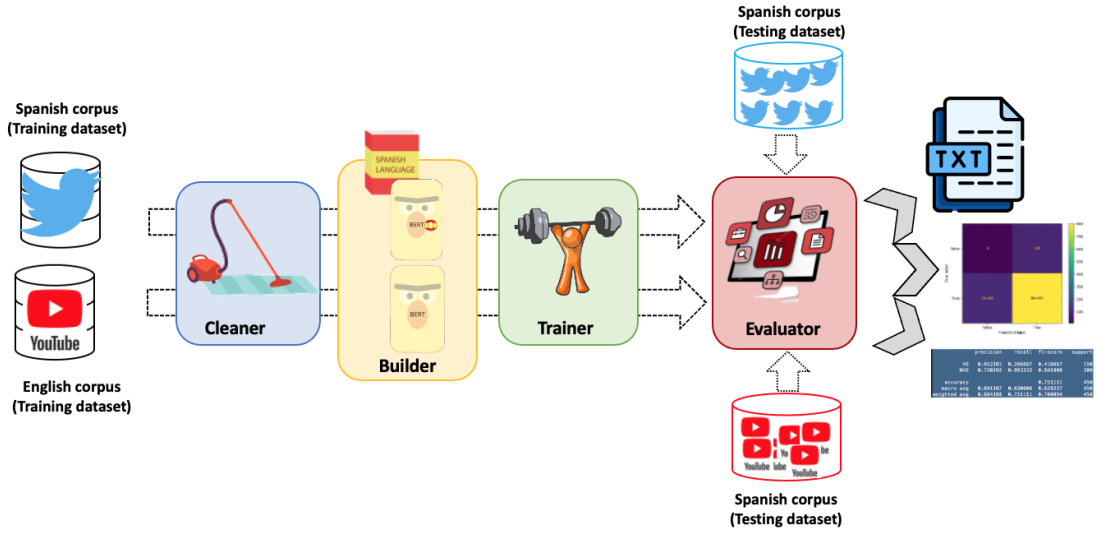

Figure 1: Architecture of the proposed system.

The built pipeline architecture is flexible since each task is independent, enabling interchange of the tools and models used in each one easily. It works as follows: first, the input data is

processed in the cleaning phase, and URLs, emojis, and stop words are removed. Second, a model is built and fine-tuned depending on the task addressed. Finally, the assessment phase is conducted, and the test data set is employed to assess its accuracy. As a result, three different outcomes can be obtained: a file with the samples classified, a picture with the resulting confusion matrix, and the evaluation metrics values.

4. Results and Discussion

The metrics selected to evaluate the performance of the participant systems are precision, recall and F1-score. Table 2 compiles the results reached for each classification task thrown in the challenge.

Table 2
Results on the official test set.

		Evaluation		
Models	label	Precision	Recall	F1-score
SpanishBETO	HS (Hope Speech)	0.89	0.32	0.47
	NHS (Non Hope Speech)	0.74	0.98	0.84
	Macro AVG	0.82	0.65	0.66
EnglishBERT	HS (Hope Speech)	0.02	0.19	0.03
	NHS (Non Hope Speech)	1	0.95	0.97
	Macro AVG	0.51	0.57	0.5

Two zones can be easily differentiated in the table, one for each experiment. At the top are the results obtained from the Spanish dataset, the one collected from Twitter. Conversely, at the bottom, the outcomes harvested by the remaining dataset, which was collected from Youtube. As can be seen, the BETO model achieves quite good results, above all for precision, where 0.89 and 0.74 in classifying tweets in HS (Hope Speech) and NHS (Non-Hope Speech), respectively. In the recall, in turn, the model behaves slightly differently, existing a notably different almost 60% between both outcomes.

After obtaining this fitful distribution, above all in the recall values, we believed this weird behaviour is due to an unusual distribution of the tweets in the datasets. The unbalanced issue was dismissed since, in Table 1, it can be seen there is not too much difference between the two sets of samples, containing 791 and 821, on HS and NHS, respectively. Hence, to dive into this hypothesis, we decided to build a word cloud for having a graphical representation of the word distributions considering both training sets. Figures 2 depict the word clouds built, where it can be seen that both sets contain similar highly used words. We believe this is one of the primary constraints responsible for the poor results obtained in recall since they do not have enough discriminating strength to assist the model in the classification task. Besides, we think this hypothesis is affirmed through the matrix confusion of Figure 3, where there is a high amount of misclassified tweets if we compare HS to NHS. If we look at the word clouds again,

(a) Wordcloud of HS tweets

(b) Wordcloud of NHS tweets

Figure 2: Graphical representation of the word frequency of the Spanish dataset delivered.

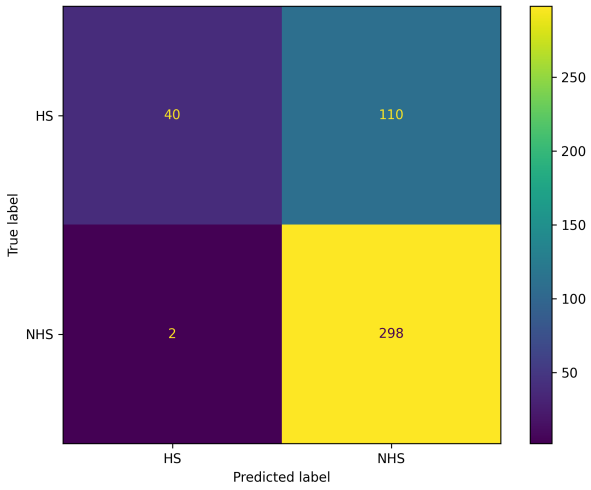

Figure 3: Confusion matrix computed from the Spanish dataset.

it can be easily differentiable several words not contained in both datasets and with a negative connotation like "*paliza*", "*homofobia*", among others. We think these words stand for these discriminating words that make the difference, increasing the models' preciseness and reducing the amount of misclassified tweets.

Considering the results obtained in the English task, there is a high difference between the HS and NHS, above 0.9 in precision and recall. In this case, we deduce three hypotheses from these wicked results. Firstly, the unbalanced dataset theory, since the NHS represent 90% of the dataset's size. However, we think this issue does not assume a clear drawback since, with reduced samples, the model in Spanish experiments is able to reach reasonable plausible outcomes. Second, it is focused on examining the word number used in comments since Youtube

comments admit more words than tweets, 10.000 characters on Youtube, against 280 characters on Twitter. Taking this hypothesis as a reply is weak since the BERT model requires less amount of words to reach better performance. Thus, the last hypothesis is specifically related to the content and is based on the premise analysed in the Spanish task. Then, we followed the same investigation procedure, conducting a frequency analysis of the comments' content and building word clouds. Figure 4 does not show very discriminating features appearing in the dataset that may help the model discern between the two sets, selecting the NHS category finally since it is the most weighted class. It is noteworthy that in the matter confusion in Figure 5, almost the whole dataset has been classified as an NHS label, and only 3% of comments escaped from this category.

(a) Wordcloud of HS tweets

(b) Wordcloud of NHS tweets

Figure 4: Graphical representation of the word frequency of the English dataset delivered.

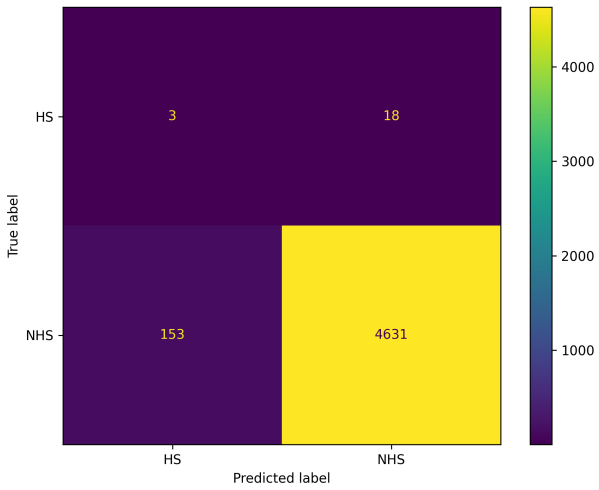

Figure 5: Confusion matrix computed from the English dataset.

Regarding the results obtained in the challenge, if we look at the leaderboard for the English recognizing task, our method got the second position with a 0.5026 score, having a difference of 0.0014 from the winner and an advantage of 0.0076 from the third competitor. Concretely, it is worth stressing we got the best precision in the Non-Hope Speech task and a raised outcome in the recall metric. On the other hand, the proposed strategies accomplished notable results on the NH classification tasks, reaching positions over the mid-table.

5. Conclusions

This article describes the proposed system for the Hope Challenge located in IberLEF 2023 shared evaluation campaign. The challenge consisted of two tasks, where Hope Speech detection had to be addressed in two different languages, Spanish and English. To face both tasks required configuring and developing two Deep Learning models. For the English language task, we employed the BERT pre-trained model. Similarly, for the Spanish, we used the Spanish version of BERT, namely, BETO. The models' performance in both tasks was notable, achieving competent results that enable our proposal to reach high positions in the leaderboard.

In future work, we think we lacked to conduct a deeper analysis of the provided dataset since we believe that some of the low results obtained are due to some aspect that was not analysed in sufficient detail. Besides, we would like to try other language models to analyse their performance and compare them with the ones submitted to the challenge.

Acknowledgments

This work has been partially supported by projects DOTT-HEALTH (PID2019-106942RB-C32, MCI/AEI/FEDER, UE), grant "Programa para la Recualificación del Sistema Universitario Español 2021-2023", and the project M2297 from call 2022 for impulse projects funded by Rey Juan Carlos University.

References

[1] A. Gowda, F. Balouchzahi, H. Shashirekha, G. Sidorov, Mucic@ lt-edi-acl2022: Hope speech detection using data re-sampling and 1d conv-lstm, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 161–166.

[2] D. Hardman, Pretending to care, Journal of Medical Ethics (2022).

[3] F. Balouchzahi, S. Butt, G. Sidorov, A. Gelbukh, Cic@ lt-edi-acl2022: Are transformers the only hope? hope speech detection for spanish and english comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 206–211.

[4] 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.

[5] 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.
[6] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, arXiv preprint arXiv:1909.12940 (2019).
[7] A. Hande, S. U. Hegde, S. Sangeetha, R. Priyadharshini, B. R. Chakravarthi, The best of both worlds: Dual channel language modeling for hope speech detection in low-resourced kannada, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 127–135.
[8] N. Ashraf, A. Rafiq, S. Butt, H. M. F. Shehzad, G. Sidorov, A. Gelbukh, Youtube based religious hate speech and extremism detection dataset with machine learning baselines, Journal of Intelligent & Fuzzy Systems 42 (2022) 4769–4777.
[9] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, Z. Žabokrtskỳ, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature communications 11 (2020) 4381.
[10] C. Yu, M. Kang, Y. Chen, J. Wu, X. Zhao, Acoustic modeling based on deep learning for low-resource speech recognition: An overview, IEEE Access 8 (2020) 163829–163843.
[11] M. Yousefi-Azar, L. Hamey, Text summarization using unsupervised deep learning, Expert Systems with Applications 68 (2017) 93–105.
[12] 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).
[13] 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.
[14] H. Nourtel, P. Champion, D. Jouvet, A. Larcher, M. Tahon, Evaluation of speaker anonymization on emotional speech, in: SPSC 2021-1st ISCA Symposium on Security and Privacy in Speech Communication, 2021, pp. 1–5.
[15] B. R. Chakravarthi, Hope speech detection in youtube comments, Social Network Analysis and Mining 12 (2022) 75.
[16] K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, IIITT@LT-EDI-EACL2021-hope speech detection: There is always hope in transformers, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 98–106. URL: https://aclanthology.org/2021.ltedi-1.13.
[17] B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, R. K. Bali, P. Buitelaar (Eds.), Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, LT-EDI@EACL 2021, Online, April 19, 2021, Association for Computational Linguistics, 2021. URL: https://www.aclweb.org/anthology/volumes/2021.ltedi-1/.
[18] N. Ghanghor, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, Iiitk@ lt-edi-eacl2021: Hope speech detection for equality, diversity, and inclusion in tamil, malayalam and english, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 197–203.

[19] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications (2023) 120078.
[20] 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.
[21] 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.
[22] 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).
[23] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 1–10.