hate-alert@LT-EDI-2023: Hope Speech Detection Using Transformer-Based Models

Mithun Das
IIT Kharagpur
mithundas@iitkgp.ac.in

Shubhankar Barman
BITS Pilani, India
contact.shubhankarbarman@gmail

Subhadeep Chatterjee
Siemens EDA, India
subhadeep.ju@gmail.com

Abstract

Social media platforms have become integral to our daily lives, facilitating instant sharing of thoughts and ideas. While these platforms often host inspiring, motivational, and positive content, the research community has recognized the significance of such messages by labeling them as "hope speech". In light of this, we delve into the detection of hope speech on social media platforms. Specifically, we explore various transformer-based model setups for the LT-EDI shared task at RANLP 2023. We observe that the performance of the models varies across languages. Overall, the finetuned m-BERT model showcases the best performance among all the models across languages. Our models secured the **first** position in Bulgarian and Hindi languages and achieved the **third** position for the Spanish language in the respective task.

1 Introduction

Social media enables people to instantly share their thoughts and connect with billions of other users. However, it has been observed that malicious users sometimes exploit social media to spread harmful speech. Consequently, significant efforts have been made to detect and mitigate hate speech. Nevertheless, social media is not solely characterized by hateful messages (Das et al., 2020). People also express their feelings and the mental stress they experience in their lives, which may arise from online harassment or workplace discrimination. Analyzing people's posting patterns can provide insights into their mental health (Tortoreto et al., 2019; Chakravarthi, 2022).

In contrast, social media posts can also be inspiring, motivational, or offer positive suggestions. For instance, individuals may share stories about overcoming life's stresses, furnishing hope and encouragement to others facing similar challenges. Such posts can particularly benefit individuals feeling down or experiencing mental stress. Recognizing the significance of these positive messages, researchers have started investigating this field, labeling such posts as "hope speech" (Chakravarthi, 2020).

Hope is commonly associated with *offering promises, support, reassurance, suggestions, or inspiration to individuals during periods of illness, stress, loneliness, and depression* (Chakravarthi, 2020). Psychologists, sociologists, and social workers affiliated with the Association of Hope have concluded that hope can serve as a valuable tool in preventing suicide or self-harm (Herrestad and Biong, 2010).

By studying the dynamics of hope speech on social media, researchers aim to understand better its impact and potential in promoting well-being and mental health. This research has the potential to contribute to the development of strategies and interventions that utilize hope speech to support individuals in difficult times, ultimately fostering a more positive and supportive social media environment Chakravarthi (2020, 2022).

Although English is the most commonly used language on social media platforms, people from various linguistic backgrounds participate and share their thoughts in local languages or dialects. As a result, there is a need for a comprehensive understanding across multiple languages. To promote research on hope speech detection, the organizers of the Hope Speech Detection for Equality, Diversity, and Inclusion (LT-EDI - RANLP 2023) (Kumaresan et al., 2023)[1] shared task has introduced the task of detecting hope speech in four languages: Bulgarian, English, Hindi, and Spanish. Since our team, hate-alert, is particularly interested in low-resource languages, we participated in the Bulgarian, Hindi, and Spanish languages.

This paper presents the methodologies we employed to identify hope speech detection, which led

[1] https://sites.google.com/view/lt-edi-2023/

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 250–256, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_038

to our team achieving first place for the Bulgarian and Hindi languages and third place for the Spanish language in the overall leaderboard standings of the shared tasks.

2 Related Work

This section explores some of the related topics around hope speech detection.

2.1 Sentiment Analysis

The task of sentiment analysis is a well-studied topic in the research community (Medhat et al., 2014). Its primary objective is to identify the sentiment or opinion expressed in the text. Sentiments are categorized into positive, negative, or neutral based on the emotions conveyed in a post. Early approaches to sentiment analysis relied on lexicon-based methods (Taboada et al., 2011), where sentiment polarity was assigned to words using predefined sentiment lexicons or dictionaries. With the advancement of deep learning techniques, models like recurrent neural networks (RNNs) (Baktha and Tripathy, 2017), convolutional neural networks (CNNs) (Ouyang et al., 2015), and Long Short-Term Memory (LSTM) (Miedema and Bhulai, 2018) networks are also being utilized for sentiment analysis. These models have shown promising results in capturing sentiment patterns within text. Furthermore, Transformer-based language models such as BERT are gaining popularity in various tasks, including sentiment analysis, due to their effectiveness and performance (Pipalia et al., 2020). Sentiment analysis has also expanded to incorporate other modalities, such as images, audio, and video (Soleymani et al., 2017). This extension has given rise to multimodal sentiment analysis, which combines information from multiple modalities to gain a deeper understanding of sentiment expressed in various media formats. By leveraging multiple modalities, more comprehensive sentiment analysis can be achieved.

2.2 Harmful Speech Detection

Harmful language detection, plays a crucial role in natural language processing (NLP) and computational linguistics. Numerous studies have examined the dissemination of hateful content on social media platforms (Das et al., 2021b, 2022c). A significant line of research focuses on detecting harmful speech by developing datasets and machine learning models similar to sentiment analysis (Pramanick et al., 2021; Chandra et al., 2021; Das et al., 2022b, 2023; Das and Mukherjee, 2023). Davidson et al. (2017) conducted a notable study where they publicly released a Twitter dataset containing thousands of labeled tweets categorized as offensive, hate speech, or neither. Earlier attempts to build classifiers for harmful speech detection utilized simple methods such as linguistic features, word n-grams, bag-of-words, and so on. Das et al. (2022a) contributed to the field by developing models specifically designed to detect abusive speech in Indic languages, showcasing the effectiveness of Transformer-based models (Vaswani et al., 2017). The utilization of Transformer-based models has proven to be highly effective in detecting harmful speech(Das et al., 2021a; Banerjee et al., 2021),. Inspired by the exceptional performance of these Transformer-based models, we also employ such models, namely mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2019), for our research.

2.3 Research on Hope Speech

Only a limited amount of work has been conducted so far on hope speech detection. Chakravarthi (2020, 2022) significantly contributed by creating the HopeEDI dataset, which consists of user-generated comments from the social media platform YouTube. The dataset comprises 28,451 comments in English, 20,198 comments in Tamil, and 10,705 comments in Malayalam. The authors also implemented several baseline approaches by utilizing the developed datasets and exploring various traditional machine-learning models. In addition, several shared tasks (Chakravarthi et al., 2021, 2022) have been organized using these datasets to encourage and facilitate research on hope speech detection within the research community.

3 Dataset Description

The Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2023) shared task at RANLP 2023 focuses on the detection of hope speech in social media through a classification problem. The main objective of this shared task is to develop methodologies for detecting hope speech in multiple languages. Table 2 presents the class distribution of the dataset, showcasing the proportions of different categories. Although our team did not participate in the English language category, we provide the dataset distribution for all languages for the sake of completeness. For the Bulgarian

Text	Label	Lang
Is topic pe aap bimar lagte hai, it's natural mr.khan ,Soch badlo	*HS*	HI
सच मे सर आपकी पढ़ाने का तरीका देख के मन करता है हमेशा हम पढ़ते ही रहे	*NHS*	HI
Наистина се забавлявах докато го гледах! Евала за това което правиш!	*HS*	BG
Estoy de acuerdo. 4 tipos de humores distintos, viva la diversidad #lgtb 🏳️‍🌈	*HS*	ES

Table 1: Example of Hope Speech. *HS*: Hope Speech, *NHS*: Not Hope Speech. Lang: Langauges

	Bulgarian		English		Hindi		Spanish	
	HS	*NHS*	*HS*	*NHS*	*HS*	*NHS*	*HS*	*NHS*
Train	223	4,448	1,562	16,630	343	2,219	691	621
Val	75	514	400	4,148	45	275	100	200
Test	150	449	21	4,784	53	268	300	247
Total	448	5,411	1,983	25,562	441	2,762	1,091	1,068
	5,859		27,545		3,203		2,159	

Table 2: Dataset statistics. *HS*: Hope Speech, *NHS*: Not Hope Speech

language, the dataset consists of a total of 5,859 data points, with 448 labeled as 'hope speech' and the remaining 5,411 as 'non-hope speech'. In the case of English, a total of 27,545 data points were shared, with 1,983 falling under the 'hope speech' category. There are 3,203 instances for Hindi, out of which 441 are labeled as 'hope speech'. Lastly, the Spanish language dataset includes 2,159 instances, with 1,091 instances categorized as 'hope speech'. One notable observation is that, except for Spanish, the other languages exhibit a significantly lower proportion of hope speech compared to non-hope speech, indicating a high-class imbalance within the datasets. We show some examples of data points in Table 1.

4 Methodology

This section discusses the preprocessing steps and various models that we implement for the detection of hope speech.

4.1 Problem formulation

We formulate the hope speech detection task in this paper as follows. Given a dataset **D** consisting of pairs ($\boldsymbol{X}$, **Y**), where $\boldsymbol{X} = \{w_1, w_2, ..., w_m\}$, represents a text sample, consisting of a sequence of words and Y represents its corresponding label, the goal is to learn a classifier $F : F(\boldsymbol{X}) \rightarrow \mathbf{Y}$, that can accurately predict the presence or absence of hope speech in unseen text samples, where $\mathbf{Y} \in \{0, 1\}$ is the ground-truth label.

4.2 Preprocessing

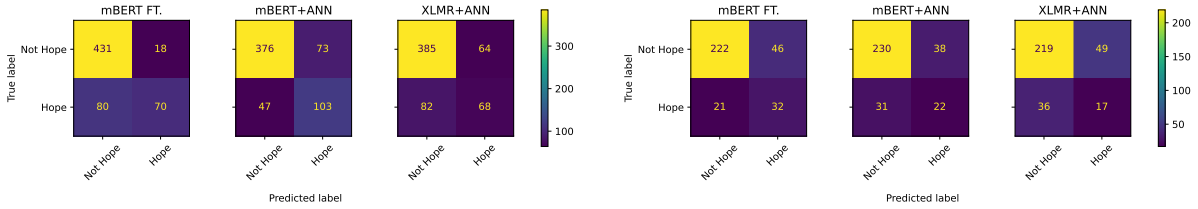
Before applying the models, we perform several preprocessing steps to prepare the data for hope speech detection. We utilize a combination of custom functions and helpful libraries such as "emoji" and "nltk" for baseline preprocessing tasks. The following pre-processing steps are performed –

- *Replacing Tagged User Names*: We replace all tagged user names with the "@user" token to remove personal identifiers from the text.
- *Removing Non-Alphanumeric Characters*: Non-alphanumeric characters, except for full stops and punctuation marks like "|" and ",", are removed. This step ensures that the machine can identify the sequence of characters accurately.
- *Removing Emojis, Flags, and Emotions*: We also remove emojis, flags, and emotional symbols from the text as they do not contribute to the semantic content of hope speech.
- *Removing URLs*: All URLs are eliminated from the text to exclude any web links that may not be relevant to hope speech detection.
- *Keeping Hashtags*: We retain hashtags in the text as they often contain contextual information that can be valuable for identifying hope speech.

By performing these preprocessing steps, we ensure that the text data is clean and optimized for the classification task.

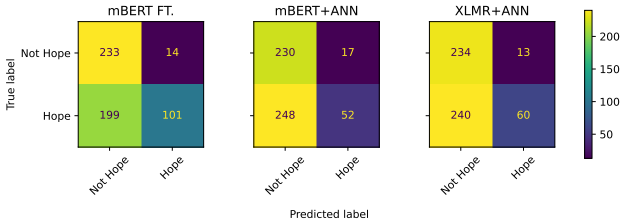
	Bulgarian			Hindi			Spanish		
Model	Acc	MF1	F1(H)	Acc	MF1	F1(H)	Acc	MF1	F1(H)
mBERT FT.	**0.836**	0.743	0.588	**0.791**	**0.678**	**0.488**	**0.610**	**0.586**	**0.486**
mBERT+ANN	0.799	**0.747**	**0.631**	0.785	0.629	0.389	0.515	0.458	0.281
XLMR+ANN	0.756	0.661	0.482	0.735	0.561	0.285	0.537	0.485	0.321

Table 3: Performance Comparisons of Each Model. FT.: fine-tuned, H: hope speech. MF1: Macro F1 Score. The best performance in each column is marked in **bold** and the second best is underlined



(a) Bulgarian

(b) Hindi



(C) Spanish

Figure 1: Confusion-matrix for all the models across languages

4.3 Models

mBERT (Devlin et al., 2019) (Multilingual BERT) represents a widely-used multilingual language model developed by Google. It utilizes the BERT (Bidirectional Encoder Representations from Transformers) architecture and has been trained on a vast corpus of text encompassing multiple languages. With its ability to comprehend texts in various languages. In our work, we employ two different approaches with the mBERT model: 1) *finetuning* and 2) *pre-trained embedding + ANN*.

For the *finetuning* process, we augment the mBERT model by adding an additional classification head and then finetune the model for the classification task. This allows us to adapt the mBERT model to capture the nuances of hope speech better. We use the `bert-base-multilingual-uncased` model[2] for our experiment.

In the case of *pre-trained embedding + ANN* setting, we pass the texts through the pre-trained mBERT model and obtain 768-dimensional feature vectors. These vectors serve as representations of the texts' semantic properties. Finally, we feed these feature vectors into an Artificial Neural Network (ANN) model to perform the classification task. This combination of pre-trained embeddings from mBERT and an ANN classifier enables us to leverage both the contextual information from mBERT and the discriminative power of the ANN for hope speech detection.

XLMR (Conneau et al., 2020) (Cross-lingual Language Model Representation) is a state-of-the-art multilingual language model developed by Facebook AI. It is built upon the Transformer architecture and trained on a vast amount of multilingual data from different languages. For the case of XLMR, we only explore the *pre-trained embedding + ANN* setting. Here again, we extract the 768-dimensional feature vectors from the `xlm-roberta-base` model[3] and feed these feature vectors into an Artificial Neural Network (ANN) model to perform the classification task.

[2] https://huggingface.co/bert-base-multilingual-uncased

[3] https://huggingface.co/xlm-roberta-base

4.4 Experimenal Setup

We finetune the mBERT model using a maximum token length of 128 and a batch size of 16. The model is trained for five epochs without utilizing early stopping. The ANN model consists of two hidden layers with 256 and 128 nodes respectively, which are then connected to an output layer with two nodes. As in this scenario, we are using pre-trained embedding, we used the maximum token length of 512 to extract features. The ANN models are trained with a batch size of 32. We run the ANN-based models for 20 epochs. For all the models, we employ the Adam optimizer with binary cross-entropy with an initial learning rate of 2e-5 and epsilon set to 1e-8. We train the models for each language separately and saved the model checkpoint for the best validation performance in terms of macro-F1 score.

5 Results

Table 3 presents the performance of each model. In the case of the Bulgarian language, we observe that although the fine-tuned mBERT model achieves the highest accuracy (0.836), the mBERT+ANN setting outperforms other models in terms of macro F1 score, achieving the highest score (0.747). For the Hindi language, mBERT demonstrates the best performance across all metrics, while mBERT+ANN achieves the second-best performance. Regarding the Spanish language, fine-tuned mBERT again attains the highest score, whereas the XLMR+ANN model becomes the second-highest scorer. Overall, we observe that fine-tuning the mBERT model end-to-end yields better scores compared to using extracted pre-trained embeddings and passing them through an ANN model. We further show the confusion matrix of each model for all the languages in Figure 1.

6 Conclusion

In this shared task, we deal with a novel challenge of detecting hope speech across multiple languages. To evaluate the performance, we employed transformer-based models such as m-BERT and XLMR. Our observations revealed that for the Bulgarian language, the **mBERT+ANN** model configuration achieved the best results. Conversely, for Hindi and Spanish, **fine-tuned mBERT** models exhibited superior performance. In the future, we plan to explore additional transformer-based models, as well as the recent LLM (Large Language Model) models, to enhance our approach in this domain further.

References

Kiran Baktha and BK Tripathy. 2017. Investigation of recurrent neural networks in the field of sentiment analysis. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pages 2047–2050. IEEE.

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. *arXiv preprint arXiv:2111.13974*.

Bharathi Raja Chakravarthi. 2020. 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*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, et al. 2022. 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*, pages 378–388.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, RL Hariharan, John Philip McCrae, Elizabeth Sherly, et al. 2021. Findings of the shared task on offensive language identification in tamil, malayalam, and kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 133–145.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *13th ACM Web Science Conference 2021*, pages 148–157.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022a. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021a. Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach. *arXiv preprint arXiv:2111.14830*.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022b. Hate speech and offensive language detection in bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 286–296.

Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *ACM SIGWEB Newsletter*, (Autumn):1–8.

Mithun Das and Animesh Mukherjee. 2023. Transfer learning for multilingual abusive meme detection. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 245–250.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021b. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022c. Hatecheckhin: Evaluating hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387.

Thomas Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Henning Herrestad and Stian Biong. 2010. Relational hopes: A study of the lived experience of hope in some patients hospitalized for intentional self-harm. *International journal of qualitative studies on health and well-being*, 5(1):4651.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Momchil Hardalov, Ivan Koychev, Preslav Nakov, Daniel García-Baena, and Kishore Kumar Ponnusamy. 2023. Overview of the third shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Fenna Miedema and Sandjai Bhulai. 2018. Sentiment analysis with long short-term memory networks. *Vrije Universiteit Amsterdam*, 1:1–17.

Xi Ouyang, Pan Zhou, Cheng Hua Li, and Lijun Liu. 2015. Sentiment analysis using convolutional neural network. In *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*, pages 2359–2364. IEEE.

Keval Pipalia, Rahul Bhadja, and Madhu Shukla. 2020. Comparative analysis of different transformer based architectures used in sentiment analysis. In *2020 9th international conference system modeling and advancement in research trends (SMART)*, pages 411–415. IEEE.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Giuliano Tortoreto, Evgeny Stepanov, Alessandra Cervone, Mateusz Dubiel, and Giuseppe Riccardi. 2019. Affective behaviour analysis of on-line user interactions: Are on-line support groups more therapeutic than twitter? In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 79–88.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.