IDIAP_TIET@LT-EDI-ACL2022 : Hope Speech Detection in Social Media using Contextualized BERT with Attention Mechanism

Deepanshu Khanna
TIET, Patiala, India
dkhanna_be19@thapar.edu

Muskaan Singh and Petr Motlicek
IDIAP Research Institute,
Martigny, Switzerland
(msingh,petr.motlicek)@idiap.ch

Abstract

With the increase of users on social media platforms, manipulating or provoking masses of people has become a piece of cake. This spread of hatred among people, which has become a loophole for freedom of speech, must be minimized. Hence, it is essential to have a system that automatically classifies the hatred content, especially on social media, to take it down. This paper presents a simple modular pipeline classifier with BERT embeddings and attention mechanism to classify hope speech content in the Hope Speech Detection shared task for Equality, Diversity, and Inclusion-ACL 2022. Our system submission ranks fourth with an F1-score of 0.84. We release our code-base here https://github.com/Deepanshu-beep/hope-speech-attention.

1 Introduction and Related Work

Social media today plays a vital role in spreading hatred and provoking people, which gives rise to hate-related crimes (Vadakkekara Suresh et al., 2021). Various hate-related terror attacks usually have a history of hate-related content in their social media accounts. Thus, large organizations such as Facebook, YouTube, Twitter are working tirelessly to detect and bring down such hateful content from their platforms (Kumaresan et al., 2021). Since hate content must not be confused with Freedom of speech and expression, thus it becomes quite challenging to reduce the number of false positives. Previously, multiple experiments have been performed for Hope Speech detection, and there are various datasets available as well for it. The top-performing model presented in the Hope speech detection shared task at the LT-EDI-2021 workshop used the XLM-RoBERTa language model (Conneau et al., 2019), a combination of the XLM and RoBERTa language model. Further, they extracted the weighted output of the final layer of the XLM-RoBERTa model using TF-IDF to filter out the error that might cause due to the mixing of various languages supported by the model. (Gundapu and Mamidi, 2021) presented a transformer-based ensembled architecture consisting of a BERT pre-trained model and a language identification model. The language identification model was used to detect if the input isn't in English. On the other hand, the BERT language model was just responsible for the binary classification of Hope Speech. (Rajput et al., 2021) presented a simple classification model which initially created the static BERT (Devlin et al., 2018) embeddings matrix of the data to extract the contextual information of the data and then experimented with various Deep Neural Networks (DNN) to train a binary classifier. Seeing the dominance of transformers to solve multiple complex applications in Natural Language Processing (NLP) became the motivation for our model. Hence, we encode the data using the contextualized BERT embeddings and train an attention network. Though it is a simple architecture with relatively few parameters, it performs efficiently and can be verified through the results.

2 Shared Task Description

For the Hope Speech Detection for Equality, Diversity, and Inclusion (HopeEDI) shared task (Chakravarthi et al., 2022), we are given YouTube comments for English, Kannada, Malayalam, Spanish and Tamil languages. Our work focuses on the English language comments in the dataset. The dataset contains 22740, 2841, and 389 comments for the training, development, and test set, respectively, annotated with labels *{Hope Speech, Not Hope Speech}* for the English database. The detailed distribution for all the languages can be seen in Table 1, and some of the examples of Hope Speech, Not Hope Speech have been shown in Table 4. Along with the release of the database, the authors also released a baseline system in which they experiment with various Machine learning al-

Label	Language-wise distribution (Train + Dev)				
	English	**Kannada**	**Malayalam**	**Spanish**	**Tamil**
Hope Speech	1962 + 272	1699 + 210	1668 + 190	491 + 161	6327 + 757
Not Hope Speech	20778 + 2569	3241 + 408	6205 + 784	499 + 169	7872 + 998

Table 1: Data distribution for the HopeEDI database.

Comment	Label
these tiktoks radiate gay chaotic energy and i love it	Hope Speech
I'm a Buddhist...! ALL LIVES MATTER...!	Hope Speech
@Paola Hernandez i never said to be intolerant and hateful..... -_-	Not Hope Speech
I say we get rid of all racist tv shows	Not Hope Speech

Table 2: Examples for Hope Speech, Not Hope Speech in the HopeEDI dataset.

gorithms such as Multinomial Naive Bayes, SVM, KNN, Logistic Regression, and Decision trees.

3 Experimental Setup

In this section, we give a detailed explanation of the experimental setup for the proposed model and depict the summarized view of it in Figure 1.

3.1 Data Preprocessing

Since the comments in raw format are highly unstructured, containing irrelevant information that may cause any AI-based model to malfunction. Hence, we preprocess the data with the following operations to convert it into a suitable understandable format.

- All of the comments are converted to lower case.
- Commonly used abbreviations such as "FYI", "ASAP", "WTF" are replaced with their original full-forms.
- Removed mentions of any users such as "@Champions" from the data.
- Updating words with additional repeated characters such as "Helloooo" are updated to their correct forms.
- Emojis in the comments are decoded and replaced with what they signify, such as ":)" is replaced with "Happy".
- All of the excessive punctuation marks are removed from the comments.

3.2 Proposed Methodology

Motivated by the efficiency of transformers in NLP, we begin by encoding the comments using the BERT language model and creating an embeddings matrix. Further, this embeddings matrix is fed to the attention network, trained to classify for Hope Speech. The architectural components of the proposed model are explained below.

3.2.1 BERT language model

Bidirectional Encoder Representations from Transformers (BERT) is a machine-learning technique based on transformers to extract contextual embeddings of unlabeled texts. Unlike static embeddings, BERT generates embeddings using bidirectional context, i.e., analyzes context from both left and right of a word. Also, BERT's attention architecture computes the attention parallelly for whole input at once, unlike other traditional models that process it sequentially.

To compute the embeddings matrix of the data, we use the RoBERTa-large-MNLI language model. Initially, the data is tokenized along with the addition of "[CLS]", "[SEP]" tokens and then encoding these tokens to get the embeddings matrix of dimensionality 768. We adopted the sliding window technique while encoding the data due to the constraint of BERT family language models to take a maximum of 512 tokens as input. Hence, we continue looping to get the embeddings until the whole data gets encoded and, finally, averaging the embeddings of these several windows to get the final embeddings of an input comment. The length of the maximum input comment is noted, and all the inputs are extended to have the same length of their embeddings by adding padding. The attention network is trained using these computed

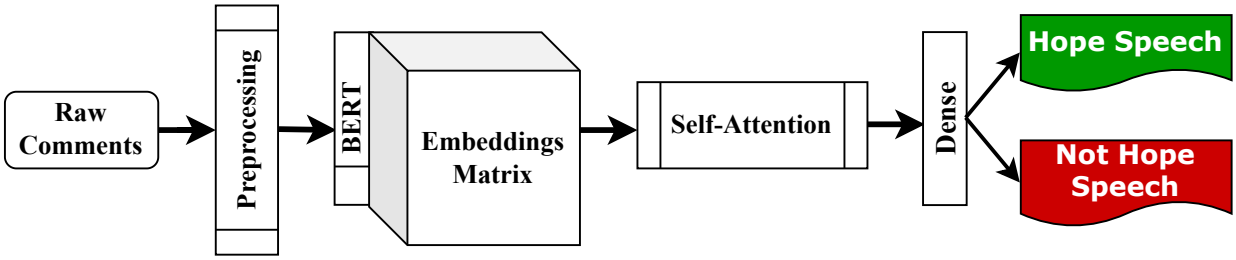

Figure 1: Flowchart illustration of the proposed model for Hope Speech classification.

	Precision		Recall		F1-Score	
Model	**Mean**	**Weighted**	**Mean**	**Weighted**	**Mean**	**Weighted**
Top performing	0.56	0.87	0.54	0.89	0.55	0.88
Proposed model	0.51	0.86	0.52	0.82	0.51	0.84
Average score	0.47	0.85	0.46	0.80	0.43	0.80

Table 3: Comparison with the top-performing model results.

Comment	**Label**
Maddona saved my Soul in 1999	Hope Speech
Her outfit is very 1990's Michael Jackson.... I like it !	Hope Speech
The way you pronounced Lewandowski gave me cancer. Levandovskee	Not Hope Speech
The end is near.	Not Hope Speech

Table 4: Examples for Hope Speech, Not Hope Speech in the HopeEDI dataset.

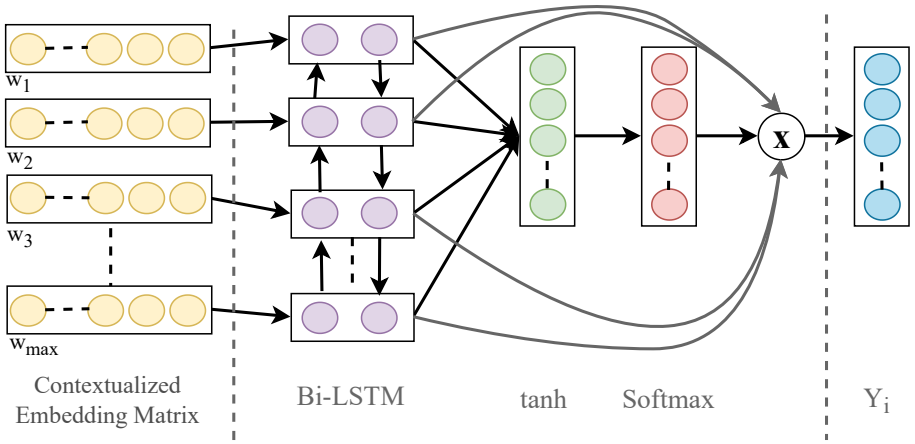

Figure 2: Architectural illustration of the attention mechanism used in the proposed model

embeddings matrices.

3.2.2 Attention mechanism

The attention mechanism has proved its efficiency by increasing accuracy in various NLP tasks. The attention module focuses on inputs having higher importance in contributing towards solving a task filtering out meaningless information, unlike in just flattening or averaging the output of convolutional layers. The architectural illustration of the attention module motivated from (Diao et al., 2020) used in our proposed pipeline is depicted in Figure 2.

Upon obtaining the word-level embeddings matrix from the pretrained RoBERTa-large-MNLI language model, we pass these obtained matrices to the Bi-LSTM layer of our attention module. The contextual representations predicted by the Bi-LSTM layer are further passed along to non-linear activation functions, namely "Tanh" and "Softmax" which can be represented mathematically as below:

$$tanh(x) = \frac{2}{1 + e^{-(2x)}} - 1 \tag{1}$$

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2}$$

Passing through non-linear activation functions updates the training weights for meaningful words accordingly. These attention weights finally help out in classifying for Hope speech.

3.2.3 Dense

The Dense layer used in a neural network is connected densely with the previous layer, i.e., each neuron of the dense layer is connected to each neuron of the last layer. The Dense layers are usually used for changing the shape of the vectors. We used the Dense layer taking input from the final layer of the attention network with the activation function "sigmoid," which can be present through the mathematical equation:

$$\sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{3}$$

3.3 Comparative Approaches explored

Additionally, we obtained the results by fine-tuning the RoBERTa-large-MNLI language model over the dataset for binary classification. With the constraints of BERT-based language models to take a maximum of 512 tokens as input, we fine-tuned the model with the first 510 tokens of the review combined with [CLS] and [SEP] token at the beginning and end of the input. We followed the same preprocessing pipeline as in our attention-based network and achieved an F1-score of 0.77 for the validation set.

4 Experiment Results and Analysis

We test our model for the English dataset of the HopeEDI database. The classification report for our proposed and the top-performing model over the test set can be seen in Table 3. The proposed model has proved itself remarkable by achieving fourth position on the leaderboard with a difference of 0.04 in F1-score from the top-performing model. For the validation set, the proposed model achieved an F1-score of 0.80 while using the same language model for binary classification achieved an F1-score of 0.77. To evaluate our results qualitatively, we performed an analysis for our prediction results in Table 4. The presented results indicate hope and non-hope speech on social media comments. Comments such as "The end is near" clearly have the potential to provoke violence, and thus can be used to encourage the masses for violence due to their negative impact on the mindset. While comments such as "Maddona saved my soul in 1999" create a positive vibe in the human attitude and give people hope and boost their confidence to support good deeds.

5 Conclusion

In this paper, we described the shared task submission on hope speech detection for English dataset. We propose a pipeline classifier architecture that uses an attention mechanism over the contextualized BERT embeddings. For future work, we intend to work upon other languages of the HopeEDI database and experiment with different neural network architectures combined with contextual embeddings.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Chinnaudayar Na-

vaneethakrishnan, John Phillip McCrae, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José Antonio García-Díaz. 2022. Findings 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*. Association for Computational Linguistics.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Yonghe Chu, Di Wu, Dongyu Zhang, and Kan Xu. 2020. Crhasum: extractive text summarization with contextualized-representation hierarchical-attention summarization network. *Neural Computing and Applications*, 32(15):11491–11503.

Sunil Gundapu and Radhika Mamidi. 2021. Autobots@LT-EDI-EACL2021: One world, one family: Hope speech detection with BERT transformer model. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 143–148, Kyiv. Association for Computational Linguistics.

Prasanna Kumar Kumaresan, Premjith, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P. McCrae. 2021. Findings of shared task on offensive language identification in tamil and malayalam. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 16–18, New York, NY, USA. Association for Computing Machinery.

Gaurav Rajput, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. Hate speech detection using static bert embeddings. In *Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15-18, 2021, Proceedings*, page 67–77, Berlin, Heidelberg. Springer-Verlag.

Gautham Vadakkekara Suresh, Bharathi Raja Chakravarthi, and John Philip McCrae. 2021. Meta-learning for offensive language detection in code-mixed texts. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 58–66, New York, NY, USA. Association for Computing Machinery.