# Autobots@LT-EDI-EACL2021: One World, One Family: Hope Speech Detection with BERT Transformer Model

**Sunil Gundapu**
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
sunil.g@research.iiit.ac.in

**Radhika Mamidi**
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
radhika.mamidi@iiit.ac.in

## Abstract

The rapid rise of online social networks like YouTube, Facebook, Twitter allows people to express their views more widely online. However, at the same time, it can lead to an increase in conflict and hatred among consumers in the form of freedom of speech. Therefore, it is essential to take a positive strengthening method to research on encouraging, positive, helping, and supportive social media content. In this paper, we describe a Transformer-based BERT model for Hope speech detection for equality, diversity, and inclusion, submitted for LT-EDI-2021 Task 2. Our model achieves a weighted averaged f1-score of 0.93 on the test set.

## 1 Introduction

With the proliferation of the Internet, the number of marginalized people looking for help and online support has grown significantly worldwide[1]. Recently, social media networks (SMN), online blogs, and online support groups (OSG) have emerged as popular online support sources. These online support networks play an essential role in marginalized people, such as individuals belonging to LGBTQIA+ (Lesbian, Gay, Bisexual, Transgender, IntersexQueer/Questioning, and Asexual people) community, people with similar health problems or disabilities, and individuals who belong to racial and ethnic minorities. And some of the research studies (Ganda, 2014) have shown that SMNs and OSGs significantly influence people's self-identification and self-understanding. So it's necessary to detect the positive content from online sources.

In this paper, we discussed the identification of Hope Speech for Equality, Diversity, and Inclusion (EDI) from the YouTube comments. Hope can be defined as a state of mind that brings fortitude, support, and reassurance to life. Hope can also come from a motivational discussion about how people deal with difficult situations and cope with them. Hope speech has a positive impact on various aspects that harm people's lives. We describe the hope speech for our task as "social media comments that give inspiration, suggestions, insight, and support". We have done various experiments on the provided Hope Speech dataset for Equality, Diversity, and Inclusion problem using different state-of-the-art machine learning and deep learning models.

## 2 Shared Task Description

We are attempting this shared task for English language only. The LT-EDI-2021 Task 2 (Chakravarthi and Muralidaran, 2021) is as follows: Given a YouTube comment or post, the system should identify its class. We have three classes 'Hope speech', 'Not hope speech' and 'Other language.' We classify the comment into the 'Hope speech' category when it promotes hope, optimism, faith, support, reassurance, suggestions, or it offers EDI values. If a comment doesn't have attributes mentioned in the hope speech class, classify it into the 'Not hope speech' class. We categorize the comment into the 'Other language' class when the comment not in the English language. Table 1 shows few sample YouTube comments from HopeEDI.

## 3 Related Work

Over the past few years, various methods have been proposed to identify hate speech on social media platforms. Alrehili (2019) surveyed the different state of the art NLP approaches to detect hate speech in OSNs, and they say there has been significant research work so far on hate speech. Considering the research work on Hope Speech, much less has been done compared to Hate Speech. By choosing the YouTube comments, Severyn et al.

[1] https://www.statista.com/topics/2409/digital-health/

*Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 143–148
April 19, 2021. ©2020 Association for Computational Linguistics

| **Example Comment** | **Label** |
|---|---|
| These tiktoks radiate gay chaotic energy and i love it | Not Hope Speech |
| I feel so base for that guy! They treated him as if he wasn't a human just because of who he loved! | Hope Speech |
| CASA. LA. FEMMEnWesT. ViLLAGEn2008nNyc | Not English |

Table 1: Example YouTube Comments

(2014) conducted a systematic study on opinion mining. For this work, the authors created a manually labeled 35K size YouTube comment dataset to model the polarity of comments based on the kernel models.

Recently, to support the Rohingyas refugees, Palakodety et al. (2019) analyze how the hope speech from the social media comments can be used to reduce tensions between India and Pakistan during the Puluma attack. In this work, the authors intend to find the hope speech in SMNs can reduce the strain and violence between two nations. The authors created a corpus of 921,235 English Youtube comments posted by 392,460 users to accomplish this work.

One of the notable research work for hope speech detection is HopeEDI (Chakravarthi, 2020) corpus creation. In this work, authors created a corpus from user-generated YouTube comments for English, Tamil, and Malayalam. Then they developed various machine learning models for benchmark results on the dataset.

## 4 Dataset Description

This paper used the corpus (HopeEDI) provided by the "Hope Speech Detection" organizers to train and tune the models. The HopeEDI for English dataset containing 28,451 samples, in that 2484 samples belongs to the 'Hope speech' class, 25,940 samples belong to the 'Not hope speech' class, and the remaining 27 samples belong to the 'Other language' class. Table 2 represents the annotated corpus distribution, and Figure 1 shows the number of comments for the class.

| **Labels(↓)** | **Train** | **Valid** | **Test** |
|---|---|---|---|
| Not Hope Speech | 20778 | 2569 | 2593 |
| Hope Speech | 1962 | 272 | 250 |
| Other Language | 22 | 2 | 3 |
| Total | 22762 | 2843 | 2846 |

Table 2: Dataset Statistics

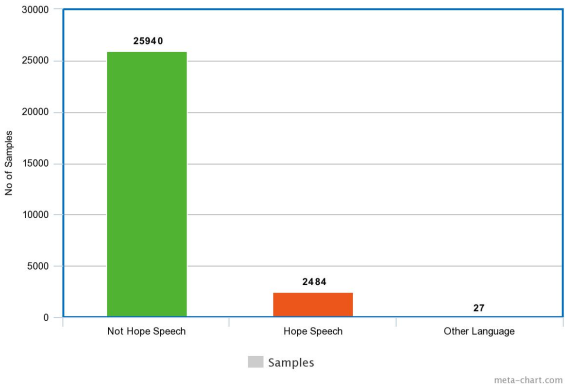

Figure 1: Distribution of samples per class

## 5 Proposed Approach

### 5.1 Data Preprocessing

Following pre-processing operations performed on the comments before feature engineering.

- All the comments in the dataset are converted to lowercase.
- Converted the online chatting abbreviations like 'ASAP', 'YOLO' into their original form by creating a slang words mapping dictionary.
- Expanded all contractions in the comments by writing regular expressions. For example, "they're" expanded into "they are", and "I'll" expand into "I will".
- In our problem, emojis play a crucial role, so we converted all the emojis into their respective keywords by using a python library called **demoji**[2].
- We removed all punctuation marks in the dataset.

### 5.2 Methodology

We started our experiments with various machine learning (ML) algorithms like Support Vector Machines (SVM), Logistic Regression (LR), Multi-Layer Perceptron (MLP), and XGBoost and deep

[2] https://pypi.org/project/demoji/

learning (DL) models like RNN (Recurrent Networks) and LSTMs (Long Short Term Memory) (Hochreiter and Schmidhuber, 1997) with various feature embeddings. However, Transformer based BERT model gave superior results to the above-mentioned techniques.

Our best hope speech detection model is ensembling of a pre-trained BERT and a rule based language identification model. The architecture of this ensemble model can be seen in Figure 2.

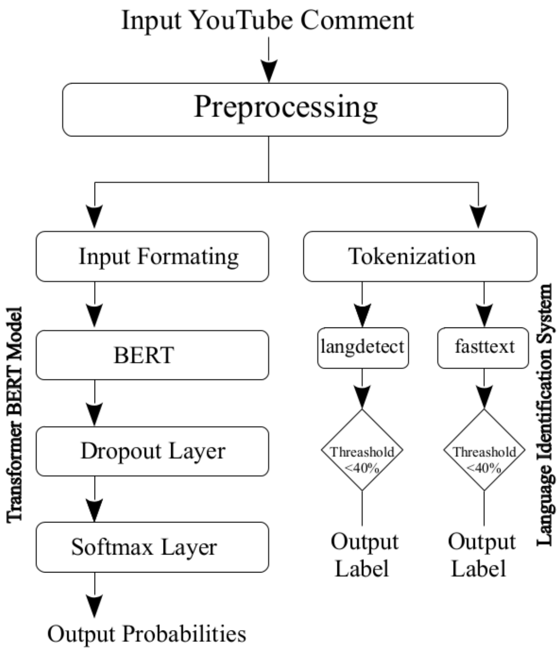

Figure 2: Ensembling Architecture

## 5.3 Transformer based BERT model

This section will describe the transformer based BERT model and then explain how we fine-tune this model to our problem.

### 5.3.1 Input Data Format

The input token sequence for the BERT model must be given in a certain format. Every input sequence must start with a [CLS] (classification token) token, and every sequence should be separated from other sequences by using a [SEP] (separation token) token. According to this BERT input data format, we added a [CLS] token to every input sequence and appended a [SEP] token after every sentence.

### 5.3.2 BERT Architecture

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a "deep bidirectional" model designed to learn deep bidirectional representations from a large unsupervised text corpus. It reads the information from both the left and the right sides of input tokens during the training phase. In our ensembling architecture, we used the BERT base case model. The base model consists of 12 transformer blocks, 768 hidden blocks, 12 self-attention heads, and a total of 110 trainable parameters. And for better context learning BERT model uses the following two separate mechanisms:

1. **Masked Language Model (MLM):** In this mechanism, 15% of the tokens in the input sequence are masked out by replacing them with a [MASK] token. Then the complete sequence is fed to a deep bidirectional Transformer model. This model tries to estimate the actual value of the masked words based on the context of unmasked tokens in the input sequence.

2. **Next Sentence Prediction (NSP):** In this task, the model learns the relationship between input sequences to distinguish the two input sequences. BERT model takes pair of sequences as input and learns to predict whether the second sequence in the pair actually follows the first sequence or it is a random sequence.

Unlike traditional directional models, which reads the input sequence either from left to right or right to left, the BERT encoder attention mechanism processes the input sequence simultaneously, allowing all input tokens in the sequence to be processed in parallel. This feature enables the model to know the context of a token based on the tokens around it. We can see all the layers of BERT architecture in Figure 3. This pre-trained BERT model fine-tuned to our problem by adding a softmax classification layer on top of the BERT model output for the [CLS] token. Below is described how an input YouTube comment goes from the BERT model and gives an output class for the input.

- First, we preprocess the input sequence then arrange the sequence according to the input data format as described earlier.

- Then the preprocessed input fed to the BERT model. The BERT model generates sequence embeddings for all the tokens in the input sequence. But we consider the [CLS] token corresponding sequence embedding to detect hope speech.

- The sequence embedding corresponding to [CLS] token passed to a softmax classification layer to get each target class's output probabilities. The softmax layer uses a softmax activation function that calculates the probability for every possible output class given in (1). To avoid overfitting, added a dropout layer before the softmax layer with the dropout of 0.2.

$$Z = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{1}$$

In the above equation, where z is the sequence embedding, Z is a vector pf softmax probabilities, and the K in the number of target classes.

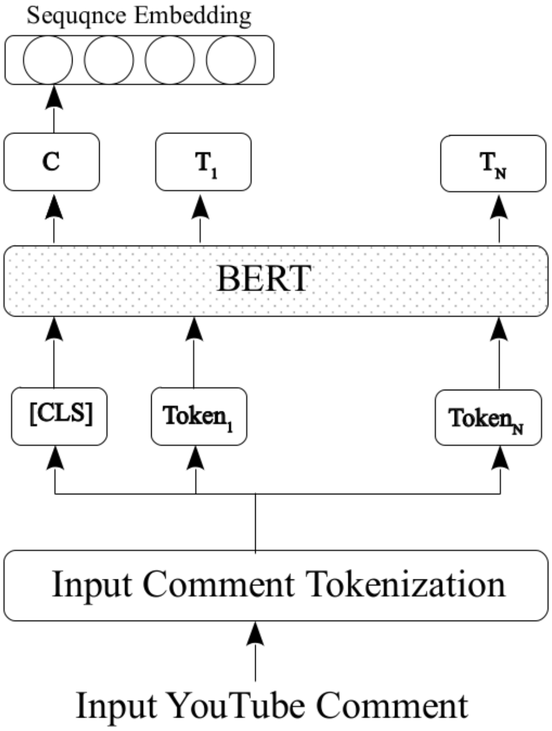

Figure 3: BERT Architecture

### 5.4 Rule based language identification model

If we examine our corpus, the 'Other language' class samples are very less (data skewing), so the statistical models are not predicting the 'Other language' samples at all. This is the main reason behind using the rule-based model in the architecture to identify those other than the English language comments. We use this model only to detect 'Other language' comments. This model identifies whether the given preprocessed comment from the English language is based on two python language identification libraries, **langdetect**[3], and **fasttext**[4]. Below explained how we identified the 'other language' class comments:

- Initially, we split the preprocessed input comment into meaningful tokens.
- Identify the language of each token in the input sequence using langdetect library. Then we calculate the percentage of English tokens in the input sequence. If the English tokens are less than 40%, then we consider that comment from the 'Other language' class—the same procedure used for the fasttext also.
- If the output of those two libraries is 'Other language,' then we predicted that the rule-based model's output is 'Other language'.

For other 'Hope speech' and 'Not hope speech' classes, we review the BERT model output only, but for the 'Other language' class, we consider the BERT and language identification model's output. And we added a constraint for only the 'Other language' class. At least any model, either BERT or Rule based model output, should be 'Other language' for the given input. If any input comment passes that constraint, we tagged a comment as 'Other language'.

## 6 Experiments and Results

To validate our approach, we conducted 3 experiments: Initially, we check the baseline model's (SVM, MLP, LR, XGBoost) results with n-gram level Term Frequency - Inverse Document Frequency (TF-IDF) vectors. Secondly, we trained the LSTM, RNN, and CNN models with pretrained Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2016) word embeddings. These models gave the better results than the baseline models but not at all contributing to the 'Other language' class whose training samples very less. Thirdly, we used the Transformer based BERT model with a rule based language identification model. This approach performed very well and gave the 0.93 wighted f1-score on test set. The rule based language identification model helped to identify the 'Other language' class samples. The

[3] https://pypi.org/project/langdetect/

[4] https://fasttext.cc/blog/2017/10/02/blog-post.html

|  | Validation Data |  |  | Test Data |  |  |
|---|---|---|---|---|---|---|
| Model | Precision | Recall | F1-Score (weighted) | Precision | Recall | F1-Score (weighted) |
| SVM | 0.91 | 0.90 | 0.91 | 0.88 | 0.87 | 0.87 |
| LR | 0.93 | 0.88 | 0.91 | 0.88 | 0.88 | 0.88 |
| MLP | 0.94 | 0.93 | 0.93 | 0.90 | 0.89 | 0.90 |
| XGBoost | 0.93 | 0.91 | 0.92 | 0.88 | 0.88 | 0.89 |
| LSTM (with fasttext) | 0.94 | 0.92 | 0.94 | 0.90 | 0.91 | 0.92 |
| CNN (with fasttext) | 0.95 | 0.91 | 0.94 | 0.91 | 0.92 | 0.92 |
| RNN (with Glove) | 0.93 | 0.92 | 0.93 | 0.91 | 0.89 | 0.90 |
| BERT with LI | 0.96 | 0.93 | 0.94 | **0.93** | **0.93** | **0.93** |

Table 3: Comparison between various model results

main motive of doing all these experiments is to assess how the proposed approach contributes to performance improvement.

We trained all our models using the training dataset and fine-tuned the model using the development. Evaluated the fine-tuned model by predicting the output labels for the unseen test set. Reported all the results in Table 3. SVM and LR performed very poorly on the test set. As already mentioned, the combination of BERT and the language recognition model has yielded better results than other approaches. And to calculate the model performance, we used the weighted average F1-score across all the classes.

As shown, Transformer based BERT model performed better than other ML and DL learning models for our hope speech detection problem. The supervised machine learning algorithms performed unsatisfactorily due to data imbalance problem. We tried to handle this class imbalance problem using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), but it did not work well. Latterly, we used a rule-based language identification model to remove the non-intended language comments from our dataset. While observing the 'Other language' class comments, we find that some of the comments are wrongly tagged. The comment's actual content is in the English language, but the annotators tagged the comment as 'Other language'. And if we observe the word cloud of the 'Other language' class in Figure 4, we can only see English words only. We note that this is also a problem for not predicting 'Other language' comments.

To find the right set of hyper-parameters, we used the validation dataset. We implemented the ML algorithms using scikit-learn[5] and DL using Keras[6]. To implement BERT model used the HuggingFace[7], it's a PyTorch based transformer library. And all details of BERT model hyper-parameters are summarized in Table 4.

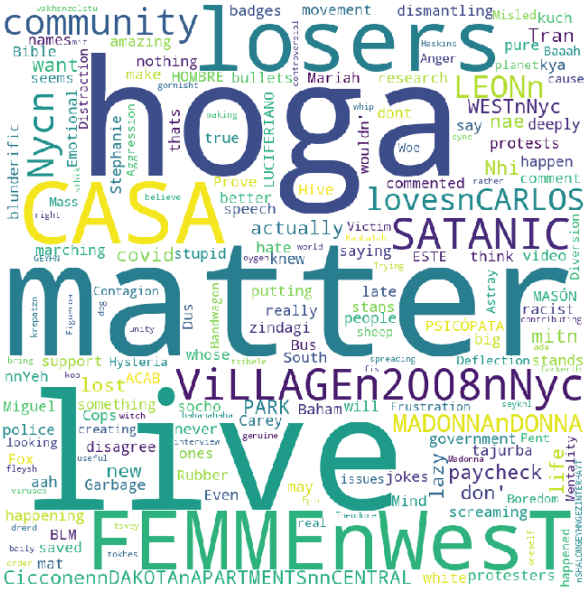

Figure 4: Word cloud of 'other language' class

| Hyper-parameters | Values |
|---|---|
| Model Type | BERT-Base |
| Learning Rate | 23-5 |
| Optimizer | Adam |
| Batch Size | 16 |
| Maximum Length | 128 |
| Epochs | 10 |

Table 4: Hyper-parameters of BERT

[5] https://scikit-learn.org/stable/

[6] https://keras.io/

[7] https://huggingface.co/transformers/model_doc/bert.html

## 7 Conclusion

This paper described a transformer-based pre-trained BERT model with a rule-based language identification system to detect hope speech in the YouTube comments. The BERT model helped for the better contextual representation of words in the comment, and the language identification model-assisted in detecting ‘Other language’ comments. In future work, we will handle the class imbalance problem efficiently by improving the dataset. And We will also explore with other transformer models like RoBERTa, XLNet, Albert, FLAIR, ELMo, etc., for a superior hope speech detection.

## References

A. Alrehili. 2019. Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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 and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357.

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

Madison Ganda. 2014. Social media and self: Influences on the formation of identity and understanding of self through social networking sites.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2019. Hope Speech Detection: A Computational Analysis of the Voice of Peace. *arXiv e-prints*, page arXiv:1909.12940.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

A. Severyn, Alessandro Moschitti, O. Uryupina, Barbara Plank, and Katja Filippova. 2014. Opinion mining on youtube. In *ACL*.