ssn_diBERTsity@LT-EDI-EACL2021:Hope Speech Detection on multilingual YouTube comments via transformer based approach

Arunima S, Akshay R, Avantika Balaji, Thenmozhi D, Senthil Kumar B
Sri Sivasubramaniya Nadar College of Engineering
{theni_d,senthil}@ssn.edu.in

Abstract

In recent times, there exists an abundance of research to classify abusive and offensive texts focusing on negative comments but only minimal research using the positive reinforcement approach. The task was aimed at classifying texts into 'Hope_speech', 'Non_hope_speech', and 'Not in language'. The datasets were provided by the LT-EDI organisers in English, Tamil, and Malayalam languages with texts sourced from YouTube comments. We trained our data using transformer models, specifically mBERT for Tamil and Malayalam and BERT for English, and achieved weighted average F1-scores of 0.46, 0.81, 0.92 for Tamil, Malayalam, and English respectively. In the task results, our approach ranked second, fourth and fourteenth place in English, Malayalam and Tamil respectively.

1 Introduction

The world today is rife with uncertainty. With the new COVID'19 strain proliferating across the globe, our chances to return to normalcy soon seem rather bleak. It is during testing times like these when people in general, crave for positive reinforcement. Even apart from this pandemic, one can argue that we cannot get enough inspiration and more often than not, employees are burdened with stress in the workplace. Moreover, there are marginalized communities, such as but not limited to, LGBTIQ (Lesbian, Gay, Bisexual, Transgender, Intersex, and Questioning), racial and ethnic minorities, which mostly rely on having faith and being hopeful for their complete acceptance in society (Puranik et al., 2021; Ghanghor et al., 2021). With the wide usage of the internet, now intensified by the ongoing pandemic, there are more people seeking the same kind of reinforcement through online forums.

YouTube, being a platform connecting billions of users across the internet, has gained an outstanding popularity across the globe (Thavareesan and Mahesan, 2019, 2020a,b). With almost 30,000 hours of content uploaded every hour, the videos from YouTube are updated showcasing the latest trends (Chakravarthi et al., 2020c,a). With the commenting facility available, one can glean a plethora of opinions from different people through this. Given the flexibility of speech in most nations in the world, this can be touted as a bane as well leading to rigorous research in Offensive Speech Detection and Hate Speech Detection (Chakravarthi et al., 2020d; Mandl et al., 2020; Chakravarthi et al., 2020b, 2021; Suryawanshi and Chakravarthi, 2021). However, there hasn't been a proportionate amount of research focusing on Hope Speech Detection. Research focusing on positive speech can have a variety of applications. With the growing concerns over mental health, recommendation engines can tailor content in a way which spreads positive comments and hope.

In our paper, we have approached Hope Speech Detection using models trained from the dataset provided by the LT-EDI organisers with the data obtained from YouTube comments in English, Tamil and Malayalam. We have tried traditional models like SVM, Logistic Regressions and Naive Bayes as well as transformers like MT5 and BERT. We achieved promising results using multilingual BERT for Tamil and Malayalam, and BERT for English YouTube comments and the task was implemented using the same.

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 92–97
April 19, 2021. ©2020 Association for Computational Linguistics

2 Related Works

The Hope Speech Dataset for Equality, Diversity and Inclusion was constructed by Chakravarthi (2020), from YouTube comments, in an attempt to encourage research on detecting positive content online. With the increasing usage of social media amongst different communities across the globe, this dataset encompasses two Indian regional languages, namely Tamil and Malayalam, along with English. The work done by Palakodety et al. (2019) focused on how Hope Speech Detection can be used to diffuse tension between politically hostile nations. One of the authors' main contributions include a novel language detection technique using Polygot word embeddings and a study analysing the temporal shifts between pro-war and pro-peace intentions of social-media users.

Recently, mining data from social media has been of interest in the field of NLP. Roy et al. (2021) has proposed a transformer based approach for Hate Speech Detection on twitter tweets for multiple languages, namely English, German and Hindi. They further classify the detected negative tweets into Hate Speech, Offensive and Profane speech. Vashistha and Zubiaga (2021) have performed a comparative study between transformer models such as BERT and XLM-RoBERTa and traditional models such as Logistic Regression to classify multilingual text in two languages, Hindi and English from online platforms, into three classes, Abusive, Hateful or Neither. The authors have also presented information regarding different model sizes and their impact on the training and inference times.

The worldwide impact of the COVID'19 virus has led to the surge of varied news articles disseminated without much scientific backing. The detection of COVID'19 fake news in English and hostile posts in Hindi has been done by Sharif et al. (2021). Various techniques such as SVM, CNN and Bi-LSTM have been employed for these tasks. In order to tackle Out-of-Vocabulary words, Dong and Huang (2018) proposed a methodology combining the pre-trained word embeddings from a large general text corpus with the ones generated from the training corpus based on word2vec. The proposed new representation contains information from both the sources and using vector concatenation, the authors were able to integrate them effectively.

3 Data Analysis

The dataset released by LT-EDI 2021 (Chakravarthi and Muralidaran, 2021) for the three languages, English, Malayalam and Tamil consisted of 28451, 10705 and 20198 YouTube comments respectively. This was distributed amongst the Training, Validation and Test datasets in the following way:

Language	Training	Validation	Test
English	22,762	2,843	2,846
Malayalam	8,564	1,070	1,071
Tamil	16,162	2,018	2,020

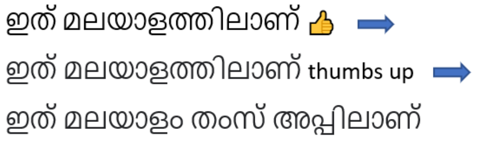
Table 1: Data distribution

Language	Hope	Non_Hope	Other Lan
English	2,484	25,940	27
Malayalam	2,052	7,765	888
Tamil	7,899	9,816	2,483

Table 2: Class-wise distribution of data

The training and the validation dataset consisted of two columns each, "Sentence" and "label". The preprocessing performed for all three languages had a few common steps including:

- Removing extra spaces, punctuations and special characters such as '@'.
- Replace the emojis with text using the demoji[1] library. The emojis were converted into text in English. For Malayalam and Tamil, these text strings were further translated into Malayalam and Tamil scripts respectively. For example,



ഇത് മലയാളത്തിലാണ് 👍 ➡

ഇത് മലയാളത്തിലാണ് thumbs up ➡

ഇത് മലയാളം തംസ് അപ്പിലാണ്

- Label encoding for the train and validation dataset.

Additionally, for the English language, we handled Out Of Vocabulary words (OOV). OOV words are words that are not present in the training dataset. These may be words that are abbreviated or misspelled, and by extension, not present in the English

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

language dictionary. These words do not have embeddings and increase the randomness in the input and can cause misclassifications. To handle this problem, we created a dictionary of length 70, containing the most frequently occurring OOV words in the train dataset and replaced them with their corrected usage. Apart from this, we also normalized contractions using a predefined list. For example, "couldn't" is replaced with "could not". Substituting these contractions makes the text standardized.

For English and Malayalam, Synthetic Minority Oversampling Technique (SMOTE (Chawla et al., 2002)) is used for oversampling the imbalanced classes present in the dataset. This is achieved by oversampling the minority class present in the dataset. SMOTE chooses data points that are nearby in the feature space, draws a line between the data points and draws a new sample at a point along that line. The model submitted for all three languages is BERT (multilingual BERT for Tamil and Malayalam). To overcome the problem of class imbalance, balanced weights were calculated using the scikit-learn[2] library and given as an argument to the model.

4 Methodology

We have employed traditional and transformer based approaches to detect the hope speech present in YouTube comments.

4.1 Traditional Learning Approach

In this approach, we use the classifiers, Support Vector Machine (Cortes and Vapnik, 1995), Naive Bayes (Rish et al., 2001) and Logistic Regression[3] to build the models for all the languages. SVM works on the principle of finding a hyperplane in an N-dimensional space, where N denotes the number of features that distinctly classifies the data points. The objective is to maximize the margin distance between the hyperplane and all data points so that classifying in the future can be done with more confidence. NB is a supervised learning algorithm based on the Bayes theorem i.e. it assumes that features are independent of each other and there is no correlation between them. It is responsible for finding the class of observation (data point) given the values of features. Logistic regression (LR) is a statistical method which finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. LR uses the log odds ratio and an iterative maximum likelihood method which makes it more appropriate for non normally distributed data.

These models are trained using the preprocessed train data and tested against the dev dataset. TF-IDF scores are used for performing word vectorization and a maximum of 5000 unique features are obtained. SMOTE is used for oversampling the imbalanced classes present in the English and Malayalam datasets. The below mentioned tables contain the validation results of the traditional learning models of the three languages.

Model	Accuracy	F1-Score
SVM	0.75	0.80
NB	0.74	0.79
LR	0.79	0.82

Table 3: English Validation Results

Model	Accuracy	F1-Score
SVM	0.82	0.80
NB	0.79	0.75
LR	0.81	0.79

Table 4: Malayalam Validation Results

Model	Accuracy	F1-Score
SVM	0.60	0.59
NB	0.60	0.59
LR	0.59	0.58

Table 5: Tamil Validation Results

Using SMOTE, we were able to obtain more balanced result, improving the recall scores for minority classes in the English and Malayalam validation set. The recall for the English language improved by two-fold upon using SMOTE and the Malayalam language saw a 53% increase in the same.

4.2 Transformer Based Approach

For this approach, we used BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) for English and mBERT(multilingual BERT) for the Tamil and

[2] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Malayalam languages. The BERT model operates on the principle of an attention mechanism to learn contextual relations between words. The transformer encoder used is bidirectional, as opposed to directional models, which reads the text input sequentially. This bidirectional behavior is responsible for allowing the model to learn the context of a word based on all of its surroundings (left and right of the word). This makes it very useful for classification tasks. To perform multilingual tasks, the mBERT model provides sentence representations for 104 languages.

We also implemented the Multilingual Text-to-Text Transfer Transformer (mT5) (Xue et al., 2020) for Tamil and Malayalam datasets. The mT5 is a multilingual variant of the "Text-to-Text Transfer Transformer" (T5) model. The mT5 model is trained on 101 languages and has achieved state-of-the-art results on a wide variety of cross-lingual NLP tasks. For both these transformer models, the preprocessed train data was given as input for training and the dev set was given for validation. In order to overcome the class imbalance problem, balanced weights were assigned to all three classes during the training of the respective models.

Model	Language	Accuracy	F1-Score
BERT	English	0.93	0.92
mBERT	Tamil	0.58	0.56
mBERT	Malayalam	0.78	0.79
mT5	Tamil	0.44	0.35
mT5	Malayalam	0.77	0.76

Table 6: Validation Results

The BERT model performed extremely well on the validation dataset for all the languages. mBERT outperformed the mT5 model for Tamil and Malayalam languages by a fair margin. The BERT and mBERT model outperformed the traditional learning models on the validation dataset based on the classwise scores. For the Tamil and Malayalam languages, there was a 38% and 27% increase in the F1-score of the minority class on the Validation datasets.

5 Results

In the LT-EDI-EACL-2021 task on Hope Speech Detection, our team, ssn_diBERTsity, achieved a top submission rank of 2nd in English, 4th in Malayalam and 14th in Tamil.

Lang	Precision	Recall	F1	Rank
Eng	0.91	0.93	0.92	2
Mal	0.82	0.81	0.81	4
Tam	0.50	0.53	0.46	14

Table 7: Weighted average precision, recall and F1-score on hope speech in English, Malayalam and Tamil text.

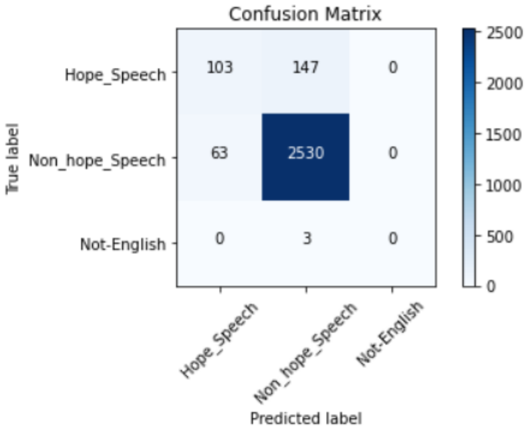


Figure 1: Confusion Matrix for English test set

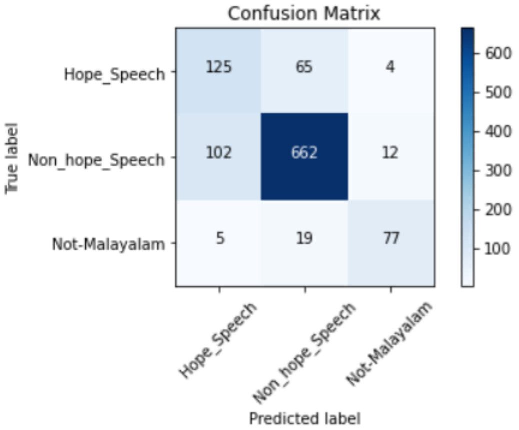


Figure 2: Confusion Matrix for Malayalam test set

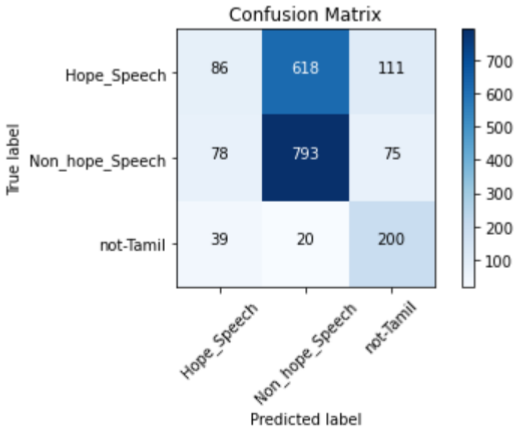


Figure 3: Confusion Matrix for Tamil test set

From Figure 1, we can conclude that out of the 2846 instances in the test set, our model for the English language has predicted 2633 true positives. From Figure 2, we can see that out of the 1071 instances in the Malayalam test set, our model has correctly predicted 864 instances. From Figure 3, we observe that out of the 2020 instances in the Tamil test set, our model has predicted 1079 true positives accurately.

Our implementation was based on the transformer model BERT for the English language and Multilingual BERT for Tamil and Malayalam. Our submission was evaluated against the test data provided by LT-EDI and achieved F1-scores of 0.92, 0.81, 0.46 for English, Malayalam and Tamil respectively.

6 Conclusion

Our implementation consists of two approaches, namely the transformer-based approach and the traditional-based approach. The model submitted for the task is BERT which outperformed the other transformer and traditional models. The validation accuracy achieved by the BERT model is 93%, 78% and 58% for English, Malayalam and Tamil respectively. BERT's capabilities extend to making more accurate predictions when dealing with newer documents even when the type of document differs significantly in key properties such as length and vocabulary. This attribute of BERT makes it the perfect choice when dealing with multiple languages and code-switching within text. In the future, we would like to extend our research in order to create a language model suitable for code-switched content and experiment with different model architectures in the hope for better performance.

References

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, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, M Anand Kumar, John Philip McCrae, Premjith B, Soman KP, and Thomas Mandl. 2020b. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India.*

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.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020c. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 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*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020d. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jianxiong Dong and Jim Huang. 2018. Enhance word representation for out-of-vocabulary on ubuntu dialogue corpus.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. 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*, Online.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2019. Kashmir: A computational analysis of the voice of peace. *CoRR*, abs/1909.12940.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. 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.

Irina Rish et al. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Neeraj Vashistha and Arkaitz Zubiaga. 2021. Online multilingual hate speech detection: experimenting with hindi and english social media. *Information*, 12(1):5.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.