CFILT IIT Bombay@LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion using Multilingual Representation from Transformers

Pankaj Singh, Prince Kumar and Pushpak Bhattacharyya
Indian Institute of Technology Bombay
Mumbai, India
pankajsingh7@iitb.ac.in, {princekumar, pb}@cse.iitb.ac.in

Abstract

With the internet becoming part and parcel of our lives, engagement in social media has increased a lot. Identifying and eliminating offensive content from social media has become of utmost priority to prevent any kind of violence. However, detecting encouraging, supportive and positive content is equally important to prevent misuse of censorship targeted to attack freedom of speech. This paper presents our system for the shared task Hope Speech Detection for Equality, Diversity, and Inclusion at LT-EDI, EACL 2021. The data for this shared task is provided in English, Tamil, and Malayalam which was collected from YouTube comments. It is a multiclass classification problem where each data instance is categorized into one of the three classes: 'Hope speech', 'Not hope speech', and 'Not in intended language'. We propose a system that employs multilingual transformer models to obtain the representation of text and classifies it into one of the three classes. We explored the use of multilingual models trained specifically for Indian languages along with generic multilingual models. Our system was ranked 2nd for English, 2nd for Malayalam, and 7th for the Tamil language in the final leader board published by organizers and obtained a weighted F1-score of 0.92, 0.84, 0.55 respectively on the hidden test dataset used for the competition. We have made our system publicly available at GitHub.

1 Introduction

The prominence of web-based media is expanding quickly because it is being used to create and share content, even by those who are ignorant of online media. Several web platforms allow users to add textual feedback on non-textual content, such as images, photos, animations, etc. With millions of videos posted by its users and billions of comments on all these videos, YouTube is undoubtedly the most famous of them.

Online social media comments/posts have been examined to identify and avoid the propagation of negativity using strategies such as detecting abusive language detection (Lee et al., 2018) and hate speech (Schmidt and Wiegand, 2017). There is a lot of work that is being done to remove the negativity from the web but Hope speech detection focuses to spread positivity by detecting content that is encouraging, positive, and supportive.

When it comes to hope speech detection there has not been much work done but recently the NLP community has started showing interest in this area. In a work by Palakodety et al. (2019), they have analyzed YouTube comments and performed the task of hope speech detection to identify hostility-diffusing content. Here the authors have not taken other aspects like equality, diversity, and inclusion into account.

Chakravarthi et al. (2020) did work in Indian languages where they manually annotated the YouTube comments for Tamil and Malayalam languages for performing sentiment analysis. In a similar work, Chakravarthi (2020) released the dataset consisting of YouTube comments with hope and not-hope speech annotation.

Hate speech is a well-researched area related to Hope speech. According to the survey done by Schmidt and Wiegand (2017), automatic hate detection was needed due to a large number of people using the net and the massive scale with which the web is growing. Zhang and Luo (2018) used deep neural networks for the task and they were able to outperform the best performing method by up to 5 percentage points in macro-average F1-score.

Multilingual BERT (Pires et al., 2019) is a variant of BERT (Devlin et al., 2019) that has been heavily used by the NLP community. Pires et al. (2019) in their work showed that multilingual rep-

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 193–196

resentation by multilingual BERT handles cross linguality without being explicitly trained for it. It also handles transfer across scripts and to code-switching fairly well. Conneau et al. (2020) proposed another variant of the BERT model called XLM-RoBERTa by pre-training multilingual models at scale. There have been various attempts to tackle problems related to Indian languages by training transformer models specifically for Indian languages. Indic BERT (Kakwani et al., 2020) and MuRIL (https://tfhub.dev/google/MuRIL/1) are two such transformer-based language models.

We plan to tackle the Hope speech discovery for the English, Tamil, and Malayalam language by obtaining representation from multilingual transformer models. Tamil and Malayalam are Dravidian languages locally spoken in the states of Tamil Nadu and Kerala respectively on Indian territory. For a country like India, where people speak many languages, code-mixing is fairly common (Barman et al., 2014; Bali et al., 2014; Gupta et al., 2018). The dataset for this task is code-mixed such as tag, inter-sentential, and intra-sentential (Chakravarthi, 2020).

The shared task was launched as a part of the first workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-EACL-2021) (Chakravarthi and Muralidaran, 2021). It was conducted for English, Tamil, and Malayalam language categories. The task was to classify a given piece of text scraped from YouTube comments into one of the three possible categories. These three possible categories or labels were:

- Label 1: Hope Speech
- Label 2: Not hope speech
- Label 3: Not in intended language

The rest of the paper is paper is organized as follows. In Section 2, we provide the dataset details and statistics. Section 3 consists of our system description and architecture details. In Section 4, we describe our experimental setups and report our results. We conclude this paper in Section 5 and briefly discuss our future plans for tackling this problem.

2 Dataset

The organizers of LT-EDI 2021 have provided the dataset for training, validation, and testing of the systems for the shared tasks. The dataset consists of pairs of text and their corresponding label. Training and validation sets were provided with labels for developing the systems for the given shared task and a test set was supplied without ground truth labels for a fair evaluation and publishing final results and team ranking among participants.However, the labels for test data were also made available once competition concluded and final results were declared. However, the labels for test data were also made available once the competition concluded and the final results were declared. The dataset was unevenly distributed across three possible classes for all three languages. Table 1 shows the class distribution of the dataset for English, Tamil, and Malayalam.

Language	Label 1	Label 2	Label 3
English	20,778	1962	22
Tamil	7,872	6,327	1,961
Malayalam	6,205	1,668	691

Table 1: Distribution of training data samples across three classes

3 System Description

The core deep learning model of our system consists of BERT-based transformer model in a multilingual setting. We did not perform any kind of pre-processing of the text data to avoid computational overhead at run time and to evaluate efficacy of multilingual transformers on raw text. We obtained a pooled 768-dimensional vector representation for the entire raw text of each instance. This vector is then fed to a softmax layer which gives the probability distribution of the sentence being from given three possible classes. This setup of trained end-to-end with cross-entropy loss function and F1-score as an evaluation metric.

4 Experiments and Results

We have experimented with various multilingual transformer-based models that were trained on multiple languages together. Multilingual models pre-trained specifically for Indian languages were also employed for the given shared tasks.

4.1 Experimental Setups

We report the performances of our system for the four multilingual transformer models. These four models are multilingual BERT, XLM-RoBERTa,

Language	mBERT	XLM-RoBERTa	IndicBERT	MuRIL
English	0.925	**0.928**	0.918	0.918
Tamil	0.570	**0.582**	0.547	0.567
Malayalam	0.850	**0.860**	0.835	0.823

Table 2: Performance (Weighted F1-score) of our system on test set for various multilingual transformer models

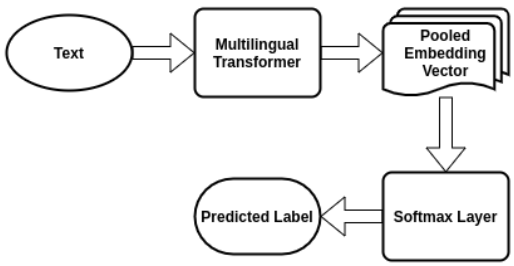

Figure 1: Architectural flowgraph of the our system

IndicBERT, and MuRIL. Among these fours, the last two models were pre-trained specifically for Indian languages. IndicBERT is an ALBERT based model that was pre-trained on 12 major Indian languages and MuRIL (Multilingual Representations for Indian Languages) is a BERT-based model pre-trained on 17 Indian languages and their transliterated counterparts.

After extensive experimentation, we finalized hyper-parameters which worked well for all four transformer models. To have a fair comparison between all models, it was also important to not have a large variation in hyper-parameters and training methodology since all the models have a common base transformer architecture as BERT. Following hyper-parameters were finally used to train the system in all four experimental setups:

- **Loss function:** Cross-Entropy
- **Optimizer:** Adam
- **Maximum token length:** 70
- **Epochs:** 20
- **Batch size:** 64
- **Learning rate:** 0.00002

4.2 Results

In this section, we present our results on the test set for the four experimental setups discussed above. In line with shared task organizers, we also used weighted F1-score as an evaluation metric.

In Table 2, the performance of our system using above mentioned four transformer model as the base is reported. This performance is on the test set. XLM-RoBERTa was the best performing multilingual model for all the three language categories. Multilingual models trained specifically for Indian languages performed at par if not better than generic multilingual models. Among the three language categories, all of our systems achieved the best scores for English and worst scores for Tamil. The high F1-score for the English language category can be attributed to the heavily skewed distribution of the dataset samples across three classes. Since Label 1 was very dominant, it increased the weighted F1-score for the English language.

We submitted our best performing system to the competition and obtained encouraging results. In the final leader board published by organizers, our team was placed at 2nd, 2nd, and 7th rank for English, Malayalam, and Tamil language categories. After extensive experimentation and hyper-parameter tuning we were able to improve our scores published on the competition leader-board for Tamil and Malayalam languages.

5 Conclusion and Future Work

In this paper, we have described our system submitted for the shared task in Hope Speech detection and reported its performance. We have extensively experimented with the possibility of employing multilingual models for given tasks. We report the performance of four different multilingual models which include transformer models specifically trained for Indian languages. Our experiments showed that transformer models pre-trained on a smaller set of languages have the potential to perform at par or better than models trained on hundreds of languages.

In the future, we plan to introduce some linguistic-based features and combined them with multilingual transformer representation to improve the overall effectiveness of the system. We also plan to systematically study the effect of abolishing the 'Not in intended language' class from the dataset as it opens up the opportunity to have a single system or deep learning model for all three

languages.

Acknowledgments

We thank the entire organizing team of LT-EDI 2021- EACL 2021 for providing us the platform and opportunity to work on such a problem of potentially high social impact.

References

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Utsab Barman, Joachim Wagner, Grzegorz Chrupała, and Jennifer Foster. 2014. DCU-UVT: Word-level language classification with code-mixed data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 127–132, Doha, Qatar. Association for Computational Linguistics.

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. 2020. 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 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 119–130.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *CoRR*, abs/1803.03662.