IRNLP_DAIICT@LT-EDI-EACL2021: Hope Speech detection in Code Mixed text using TF-IDF Char N-grams and MuRIL

Bhargav Dave
DA-IICT / Gandhinagar, India
bhargavdave1@gmail.com

Shripad Bhat
DA-IICT / Gandhinagar, India
shripadbhat30@gmail.com

Prasenjit Majumder
DA-IICT / Gandhinagar, India
prasenjit.majumder@gmail.com

Abstract

This paper presents the participation of the IRNLP_DAIICT team from Information Retrieval and Natural Language Processing lab at DA-IICT, India in LT-EDI@EACL2021 Hope Speech Detection task. The aim of this shared task is to identify hope speech from a code-mixed data-set of YouTube comments. The task is to classify comments into Hope Speech, Non Hope speech or Not in language, for three languages: English, Malayalam-English and Tamil-English. We use TF-IDF character n-grams and pretrained MuRIL embeddings for text representation and Logistic Regression and Linear SVM for classification. Our best approach achieved second, eighth and fifth rank with weighted F1 score of 0.92, 0.75 and 0.57 in English, Malayalam-English and Tamil-English on test dataset respectively. Our code is publicly available here[1].

1 Introduction

Hope can defined as a belief that the current situation would change for the better. It is vital for everyone since it helps in continuing our efforts even in difficult and unfavourable circumstances. It also encourages us to continuously improve our lives by thinking of a better future and take actions to achieve it.

In recent years social media has become one of the important aspect of our lives. People convey their opinions openly on social media on various topics. Understanding and analysing these opinions through Natural Language Processing techniques has been an active research area. Offensive content detection and classification on social media has been studied extensively. Hence, research should also focus on identifying positive online content that is encouraging and supportive. Thus identifying Hope speech in social media is an important task to gauge the opinion of people in tough times like COVID-19.

The goal of this shared task is to identify hope speech from a code-mixed dataset of comments of Dravidian Languages collected from YouTube. Shared task was introduced as three class classification of the YouTube comments into Hope Speech, Non Hope speech or Not in language, for three languages: English, Malayalam-English and Tamil-English. Shared task organizers define Hope speech as a text that offers support, reassurance, suggestions, inspiration and insight specifically in YouTube comments. The shared task focuses on hope speech for women in STEM, LGBTIQ individuals, racial minorities or people with disabilities in general for equality, diversity and inclusion (Chakravarthi and Muralidaran, 2021).

Our approaches consists of TF-IDF character n-grams and MuRIL embeddings for the text representation and Logistic Regression and Linear SVM for classification.

The remainder of this paper is organized as follows: the next section includes related work followed by Section 4 which describes the shared task dataset and Methods are presented in Section 4. Results and Analysis is given in final Section 5 and Section 6 present Conclusion.

2 Related Work

Social media content analysis is an active research area with tasks like Hate speech detection, Offensive language detection, etc. Some of the shared tasks organised recently for these are HASOC Track at FIRE[2] 2019, 2020 (Mandl et al., 2019, 2020) and OffensEval 2019, 2020 (Zampieri et al., 2019, 2020). Most popular methods of OffensE-

[1] https://github.com/bhargav25dave1996/IRNLP_DAIICT_LT-EDI-EACL2021

[2] http://fire.irsi.res.in

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 114–117
April 19, 2021.

Label	English			Malayalam			Tamil		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Not-Hope	20778	2569	2593	6205	784	776	7872	998	946
Hope Speech	1962	272	250	1668	190	194	6327	757	815
Not-in-language	22	2	3	691	96	101	1961	263	259
Total	22762	2843	2846	8564	1070	1071	16160	2018	2020

Table 1: Hope Speech Detection shared task Dataset Statistics

val (Zampieri et al., 2020) were pretrained embeddings like BERT (Devlin et al., 2019), ROBERTa (Liu et al., 2019) and ELMo (Peters et al., 2018). Best performing methods in HASOC (Mandl et al., 2020) Track use multilingual transformer based methods like XLM-ROBERTa, mBERT, etc and fine tune them for the task. TF-IDF along with Character n-grams and machine learning classifiers like Logistic Regression, SVM and XGboost also performed well and equivalent to deep learning classifiers in some of the tasks.

Hope speech detection shared task focuses on positive instead of negative comments/posts in social media. Hope speech has been to proven to be useful (Herrestad and Biong, 2010) for saving people from self harm and suicide. It has inspired people to demand rights for equality, diversity and inclusion (Chakravarthi, 2020).

3 Dataset

Hope Speech Detection shared task organizers provide datasets in three languages English, Malayalam-English and Tamil-English (Chakravarthi and Muralidaran, 2021). Dataset has been curated from Youtube comments that have been collected are pertaining to women in STEM, LGBTQ, COVID-19 and Black Lives Matters topics, using the YouTube Comment Scraper [3]. It contains 28451 comments in English, 20198 in Tamil and 10705 in Malayalam. Full statistic of dataset given in Table 1.

4 Methods

For all the YouTube comments we first preprocess the text and then create a text representation and finally classify the text using the machine learning classifiers. Figure 1 illustrates the set of steps used to classify the YouTube comments.

[3] https://github.com/philbot9/youtube-comment-scraper

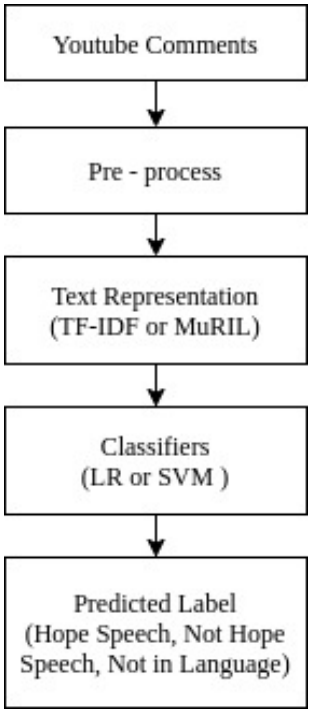

Figure 1: Steps involved in classification of Hope Speech

4.1 Pre-Processing

Since there is a lot of noise in the social media text we perform the following preprocessing operations. URL's, user mentions of the form @user, emojis, digits and punctuations is removed from the text. For English dataset, stopwords are removed and the text is lowercased.

4.2 Text representation and classifiers

Representation of the text is one of the fundamental tasks in Natural language processing. We explore two representation techniques: TF-IDF and MuRIL. TF-IDF is a very popular text representation technique which takes into account the frequency of the word in a given document and the number of documents in which a word is present. We employ Scikit-learn TF-IDF vectorizer API for obtaining the text representation. Instead of word based TF-IDF representation we make use of char-

acter n-grams based TF-IDF representation (TF-IDF (Char)) so as to effectively capture morphological variations of the words.

MuRIL[4] (Multilingual Representations for Indian Languages) is a transformer based language model trained on 17 Indian languages on self-supervised masked language modeling task. MuRIL training consists of translation and transliteration segment pairs in addition to the standard training used in Multilingual BERT. Pretrained MuRIL model is used to obtain the text representation in the form vectors of 768 dimension.

Logistic Regression (LR) and Linear SVM classifiers have been used to perform the classification of the text. The choice of the classifiers is based on the fact that they are simple, computationally inexpensive and interpretable. Scikit-learn API has been used to implement the classification task. So the four approaches we have used are TF-IDF (Char) + LR, TF-IDF(Char) + SVM, MuRIL + LR and MuRIL + SVM for each language.

5 Results and Analysis

The submission evaluation on the test data of all the three languages is shown in tables 2, 3, 4. It can be seen from the tables that our best method have been able to beat the baseline in all the three languages. Our methods achieve second, eighth and fifth rank in English, Malayalam and Tamil respectively. For English and Tamil language, TF-IDF (Char) + LR achieves the best results with weighted F1 score of 0.92 and 0.57 respectively. For the Malayalam language task, TF-IDF (Char) + SVM achieves the best weighted F1 score of 0.75.

Model	**W-Avg F1-score**
Baseline	0.90
TF-IDF (Char) + LR	**0.92**
TF-IDF (Char) + SVM	0.90
MuRIL + LR	0.87
MuRIL + SVM	0.87

Table 2: Results for the English on test dataset.

It can be observed that MuRIL is not performing good particularly in case of Tamil language. In all the languages TF-IDF character n-grams representation performed better than MuRIL. Although the training data for Tamil is higher and also balanced as compared to Malayalam, the weighted F1 score obtained for Malayalam is better than Tamil. Hence a larger dataset might not always give better results.

[4] https://tfhub.dev/google/MuRIL/1

Model	**W-Avg F1-score**
Baseline	0.73
TF-IDF (Char) + LR	0.72
TF-IDF(Char) + SVM	**0.75**
MuRIL + LR	0.61
MuRIL + SVM	0.61

Table 3: Results for the Malayalam on test dataset.

Model	W-Avg F1-score
Baseline	0.56
TF-IDF (Char) + LR	**0.57**
TF-IDF(Char) + SVM	0.56
MuRIL + LR	0.30
MuRIL + SVM	0.30

Table 4: Results for the Tamil on test dataset.

6 Conclusion and Future work

The details of our submission in Hope speech detection task have been presented in the paper. By exploring TF-IDF character n-grams and MuRIL to represent the text, we conclude that the former method is consistently performing better in all the three languages. The weighted F1 score for Tamil language is relatively lesser as compared to English and Malayalam which leads us to conclude that it is a difficult to identify in Hope speech in Tamil.

We have not explored deep learning based architectures like CNN, LSTM and Transformers for classification of Hope speech. These approaches could be a future direction for the researchers to improve the results.

Acknowledgments

We would like to thank the Hope Speech detection task organizers for giving us the opportunity to work on a new task. We also like to acknowledge the IRNLP lab at DAIICT, Gandhinagar for their contribution.

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 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.

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.

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. PMID: 20640026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.