dhivya-hope-detection@LT-EDI-EACL2021: Multilingual Hope Speech Detection for Code-mixed and Transliterated Texts

Dhivya Chinnappa
Thomson Reuters
dhivya.infant@gmail.com

Abstract

In this paper describe the shared task on hope speech detection. We present a unified framework to predict hope speech in the English, Tamil, and Malayalam datasets. Our mechanism follows a two phase approach to detect hope speech. In the first phase we build a classifier to identify the language of the text. In the second phase, we build a classifier to detect *hope speech*, *non hope speech* or *not lang* labels. Experimental results show that hope speech detection is challenging and there is scope for improvement.

1 Introduction

Artificial Intelligence models are criticized for their bias against the protected classes (Rudinger et al., 2017; Davidson et al., 2019). These biases are shown to arise from data or the model itself. There are several efforts taken to mitigate bias from the data and model perspectives (Park et al., 2018; Bender and Friedman, 2018; Mitchell et al., 2019) . Mozafari et al. (2020) present a bias alleviation mechanism evaluating the performance following a cross-domain approach.

Hope speech detection is the task of automatically detecting web content that may play a positive role in diffusing hostility on social media triggered by heightened political tensions during a conflict (Palakodety et al., 2020). We hypothesize that hope speech detection datasets could be used in evaluating the aforementioned bias alleviation mechanisms. These datasets and mechanisms could help in building AI systems that are diverse and inclusive. Additionally, with divisiveness spread across social media platforms, identifying hope speech and enhancing them would help mitigate divisiveness and animosity.

English	
God accepts everyone.	Hope
the tech industry is tough for everyone.	Non hope
CASA. LA. FEMMEnWesT.	Not Eng.
Tamil	
G..... semester exam ah pathi video uploade pannunga	Hope
Background மொக்கையா iruku	Non hope
love u mg bye.. bye... take care!!	Not Tam.
Malayalam	
Surya good. കഴിഞ്ഞ കാലം ഓക്കുന്നതാണ് ഏറ്റവും വലിയ കാര്യം. God bless u	Hope
കമ്മി ആണല്ലേ??	Non hope
I am made in India.	Not Mal.

Table 1: Examples of hope speech, non hope speech, and not *lang* in English, Tamil and Malayalam.

In this paper, we describe our approach on the hope detection shared task. We work with a multilingual corpora aiming to detect hope speech. We follow a two phase approach to accomplish the task. In the first phase, we identify the language of the input text. In the second phase we classify if the text is *hope speech* or not. We begin with describing the corpora, then analyze the datasets, explain the experimental setup, and finally discuss the results.

2 Data

We work with the hope speech detection corpora from Chakravarthi (2020). Unlike most

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 73–78
April 19, 2021.

Figure 1: Wordcloud generated from the hope speech instances of the English (left), Tamil (center), and Malayalam (right) datasets.

Figure 2: Wordcloud generated from the non hope speech instances of the English (left), Tamil (center), and Malayalam (right) datasets. Note that the Tamil wordclouds consistently have English words.

other corpora that target English texts, this corpora focuses on diversity and inclusion including two dravidian languages apart from English. The corpora contains hope speech detection datasets in three languages (i) English, (ii) Tamil, and (iii) Malayalam.

The datasets are generated from YouTube comments, and manually labeled for the three labels *hope speech*, *non hope speech*, and *not lang*. The *hope speech* label and the *non hope speech* label are self-explanatory. The *not lang* label indicates that the YouTube comment does not belong to the specific language. That is, the datasets include *not English*, *not Tamil*, and *not Malayalam* instances depending on the language of the dataset they belong. This label becomes important as the Tamil and Malayalam dataset instances are generated by social media users who are usually bilingual (English and their mother tongue). Throughout this paper, we use the label *not lang* to refer that the text does not belong to the specific language. Table 1 presents examples of *hope speech*, *non hope speech*, and *not lang* from the three datasets.

3 Analysis

The English dataset includes 28,451 instances (*hope*: 2,484; *non hope*: 25,940; *not lang*: 27), the Tamil dataset includes 20,198 instances (*hope*: 7,899; *non hope*: 9,816; *not lang*: 2,483), and the Malayalam dataset includes 10,705 instances (*hope*: 2,052; *non hope*: 7,765; *not lang*: 888). Find more about the statistics of the corpora in the original paper (Chakravarthi, 2020).

Analyzing the datasets reveal that there are several Tamil and Malayalam instances that include code-mixed or English transliterated texts. For instance, the *non hope* example for Tamil dataset in Table 1 shows a combination of code-mixed (English and Tamil) and English transliteration of Tamil. It is common for bilingual speakers around the Indian subcontinent to use several English words when uttering a sentence in their mother tongue. Additionally, as several devices did not provide easy non-English typing in the beginning of the smart phone era, most users adapted to type English transliteration of non-English sentences. This phenomena is profoundly reflected in the Tamil and the Malayalam hope speech detection datasets. In case of the English dataset, this phenomena is uncommon and there was only 0.1% of *not lang* labels. We attribute these labels to the error caused by the language detector.

Figure 1 presents wordclouds for the *hope speech* instances in the three datasets English, Tamil, and Malayalam. As we can see, the English hope wordcloud (left) has hopeful words like good, teach, etc. Interstingly, the Tamil wordcloud (center) is filled with English words and Tamil words translitered to English like *pathukonga*, *neenga*, etc. Tamil social media users tend to use code-mixed (Tamil and English) and English transliterations of Tamil texts, causing this behavior. In case of Malay-

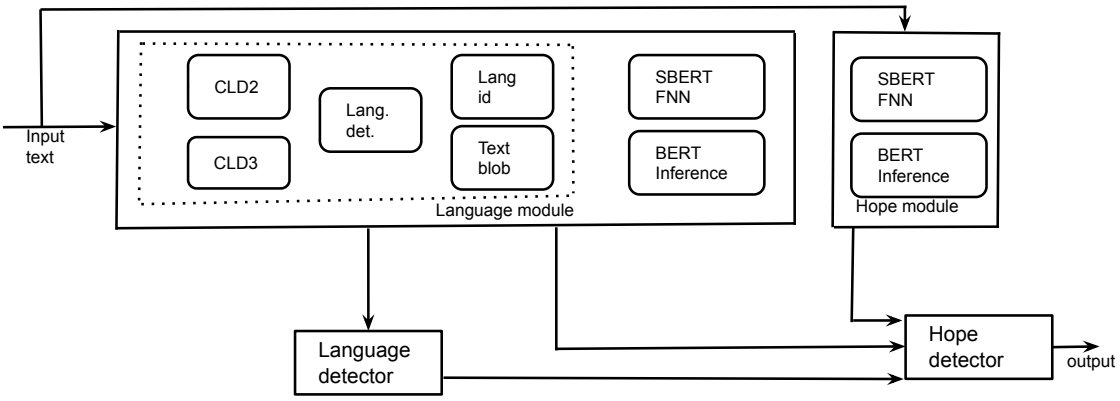

Figure 3: Architecture diagram describing the two phase hope detection process. The language detector identifies the language of the model, and the hope detector classifies the text into *hope speech*, *non hope speech*, or *not lang*.

alam, the wordcloud(right) includes both English and Malayalam words.

Figure 2 presents wordclouds for the *non hope speech* instances in the three datasets English, Tamil, and Malayalam. Unsurprisingly, the wordclouds for non hope speech instances include words with negative connotations. It is to be noted that Tamil social media users have more English influence than Malayalam social media users, despite both languages being Dravidian.

4 Experiments

As the number of *not lang* labels are considerably less than (English: 0.1%, Tamil: 12%, Malayalam: 11%) the other two labels, we specifically target to identify *not lang* labels by building a dedicated classifier. Thus we follow a two-phase approach to detect hope speech. We argue a two-phase language identification approach might be helpful as the datasets include a *not lang* label despite corresponding to a specific language. In phase 1, we identify the language of the text using a language detector. In phase 2, we use the results from phase 1 in addition to other features, to identify hope speech or not using a hope detector. The architecture is described in Figure 1.

4.1 Language detection

In this phase, we convert all *hope speech* and *not-hope speech* labels to *lang* labels, and keep the *not lang* labels as they are. Thus, we build a binary classifier using a feedforward neural network (FNN) to predict *lang* or *not lang*. We call this FNN as the language detector. The language detector takes as input (i) outputs from five language models (ii) probabilities from a vanilla feedforward network classifying *lang* and *not lang*with SBERT (Reimers and Gurevych, 2019) inputs, (iii) BERT (Devlin et al., 2019) inferences for *lang* and *not lang*.

Language models. We use five language models that take text as input and return the language of the text. The language detectors used are Compact Language Detector 2(CLD2, 2015), Compact Language Detector 3(CLD3, 2020), langid (Lui and Baldwin, 2012), textblob language detector (Loria, 2018), and langdetect (Nakatani, 2010). We use multiple language models rather than one language model improve the performance of the language detector. We follow the same approach for all language datasets including English, Tamil, and Malayalam.

SBERT FNN. We generate SBERT embeddings for each input text and feed it to a vanilla feedforward network that predicts *lang* or *not lang*. The output probabilities are passed to the language detector FNN. For the English dataset we use the *bert-base-nli-stsb-mean-tokens* model (Reimers and Gurevych, 2019) to generate SBERT embeddings. For Tamil and Malayalam datasets, we use the *distiluse-base-multilingual-cased* model (Sanh et al., 2019).

BERT inference. Here, we fine tune BERT models to predict *lang* or *not lang* obtaining BERT inferences. We hypothesize that hate speech models are useful in identifying hope speech detection, and use hate speech

	English			Tamil			Malayalam		
	P	R	F	P	R	F	P	R	F
Hope speech	0.56	0.56	0.56	0.56	0.52	0.54	0.63	0.46	0.54
Non hope speech	0.96	0.96	0.96	0.61	0.66	0.63	0.85	0.92	0.88
Not lang	0.00	0.00	0.00	0.61	0.54	0.58	0.83	0.66	0.74
W. avg.	0.92	0.92	0.92	0.59	0.59	0.59	0.81	0.82	0.81

Table 2: Results obtained with hope speech identified from text (YouTube comments) across the three languages English, Tamil, and Malayalam.

BERT models if present for a specific language. For English, we use the BERT models *dehatebert-mono-english* (Aluru et al., 2020) and *twitter-roberta-base-hate* (Barbieri et al., 2020). For Tamil, we use the BERT models *tamillion* (Doiron, 2020) and *bert-base-multilingual-uncased* (Devlin et al., 2018). For Malayalam we use the BERT model *bert-base-multilingual-uncased* (Devlin et al., 2018). Thus we have two BERT inferences results for English (*dehatebert-mono-english, twitter-roberta-base-hate*) and Tamil (*tamillion, bert-base-multilingual-uncased*), and one result for Malayalam (*bert-base-multilingual-uncased*).

The outputs from language models, probabilities from the SBERT vanilla FNN, and the BERT inferences make up the language module. The outputs from the language module is fed as an input to the language detector and the hope detector.

4.2 Hope detection

In this phase we predict the labels *hope speech*, *non hope speech*, or *not lang* using the hope detector. Similar to the language detector, the hope detector is a FNN that takes as input (i) outputs from the language module, (ii) outputs from the hope module, and (iii) probabilities from language detector.

Outputs from the language module. The same ouputs from the language module as described in 4.1 is given as the input to the hope detector.

SBERT FNN. This is similar to the SBERT FNN described in 4.1 except that it predicts *hope speech*, *non hope speech*, or *not lang*.

BERT inference. This is also very similar to the SBERT inferences described in 4.1, except that the BERT models are finetuned to predict *hope speech*, *non hope speech*, or *not lang*.

This **SBERT FNN** and **BERT inference** make up the hope module. We note that the probabilities from the **SBERT FNN** and **BERT inference** for language module and the hope module are different, as they were trained with different labels (language labels vs. hope labels).

5 Results

We present test results for English, Tamil, and Malayalam datasets in Table 2. Regarding the English dataset, the *non hope speech* labels achieve higher performance than *hope speech* labels (F1: .95 vs. .56). None of the *not lang* labels are predicted correctly. This poor performance for the *not lang* labels can be attributed to the imbalanced label distribution. There were only 3 *not lang* labels in the test set.

Regarding the Tamil dataset, the performance on all the three labels are comparable (F1: .52 vs. .66 vs. .54). Note that the Tamil dataset includes several code-mixed and transliterated texts. This phenomena can be attributed to why the classifier struggles in identifying the correct label.

Regarding the Malayalam dataset the performance of the label *hope speech* is worse than the others. Additionally, it is relatively easier to predict *not lang* labels in Malayalam.

The experimental results show detecting *hope speech* is difficult regardless of the language. Even in the English dataset where there are no transliterations or code-mixing, the classifier struggles. While we infer that hope detection is a difficult task, code-mixing and transliterations in Tamil and Malayalam increases the complexity of the problem. Chakravarthi and Muralidaran (2021) describe the results and techniques from the other participants of the hope speech detection shared task.

6 Conclusion

This paper describes the shared task on hope speech detection It targets to detect hope speech from a multilingual corpora. The corpora includes datasets in three languages English, Tamil, and Malayalam. First we conduct an analysis over the corpora finding code-mix and transliterated texts in the Tamil and Malayalam datasets. Next, we build a two phase mechanism to identify hope speech. In the first phase we detect the language of the text. In the next phase, we classify the text into *hope speech*, *non hope speech*, or *not lang*. Finally, we discuss the results concluding that hope detection is a challenging task.

References

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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.

CLD2. 2015. Compact language detector 2. https://github.com/CLD2Owners/cld2.

CLD3. 2020. Compact language detector 3. https://github.com/google/cld3.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

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.

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.

Nick Doiron. 2020. Tamillion bert. https://huggingface.co/monsoon-nlp/tamillion.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model.

Shuyo Nakatani. 2010. Language detection library for java.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.