Hopeful NLP@LT-EDI-EACL2021: Finding Hope in YouTube Comment Section

Vasudev Awatramani
Dept. of Computer Science and Engineering
Maharaja Agrasen Institute of Technology
New Delhi
vasudev.w13@gmail.com

Abstract

The proliferation of Hate Speech and misinformation in social media is fast becoming a menace to society. In compliment, the dissemination of hate-diffusing, promising and anti-oppressive messages become a unique alternative. Unfortunately, due to its complex nature as well as the relatively limited manifestation in comparison to hostile and neutral content, the identification of Hope Speech becomes a challenge. This work revolves around the detection of Hope Speech in Youtube comments, for the *Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion* (Chakravarthi and Muralidaran, 2021). We achieve an f-score of 0.93, ranking 1st on the leaderboard for English comments.

1 Introduction

With the rampant adoption of social media, problems like hostile speech detection have caught extensive attention in Natural Language Processing (NLP) research (Chakravarthi et al., 2020; Mandl et al., 2020; Chakravarthi et al., 2021; Suryawanshi and Chakravarthi, 2021). However, there has been little work on the identification of text that promotes positivity and social well-being. As social media becomes predominant in the daily lives of people, it is crucial, not only to protect users from hateful and discriminative content but also encourage communication that triggers optimism and hope. Such expression in a narrow sense may be referred to as Hope Speech and its identification in the digital space as Hope Speech Detection (Puranik et al., 2021; Ghanghor et al., 2021). However, such a task is further challenging as the definition is highly subjective and evolving. Neutral or Positive content with no indication of hostility is not necessarily a sufficient determinant. Advocacy of ideas that promote social well-being, ethics, equality, inclusion, tolerance, diversity, a fair representation of minorities, or either the appreciation or motivation for an individual or a group, are a few indicators of Hope Speech. Moreover, criticism of oppressive or malicious elements of society may also fall under such a category. Furthermore, one does not need to express such beliefs with the present but may incorporate past or future developments.

2 Related Work

Hope Speech Detection is a nascent research task by (Palakodety et al., 2020). The authors proposed automatic identification of positive web content that may diffuse hostility on social media platforms due to political tensions around the *2019 Pulwama Terror Attack*[1]. The authors mined a multilingual (Hindi and English) corpus by scrapping comments from Youtube videos related to the crisis. The authors developed a comprehensive system that took statistical NLP features: n-grams, along with temporal sentiment scores. The system also employed language identification using polyglot FastText (Bojanowski et al., 2017), to achieve an F1-score of 78.51 and 95.48 AUC. However, the work differs from HopeEDI (Chakravarthi, 2020) as it focuses on alleviation of tension and violence and ignores other aspects of hope.

The study focuses on HopeEDI dataset as part of the *Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion*. We compare our experimental outcomes with the results with (Chakravarthi, 2020) as the baseline. Chakravarthi (2020) employed TFIdf: token frequency-inverse document frequency along with classifiers such as Multinomial Naive Bayes, K-nearest neighbours, Support Vector Machine, Decision Tree and Logistic Regression. Decision tree delivered the highest F-Score for English and Malayalam while Tamil performed well with Logistic Regression. A more thorough comparison occurs in Section 4.

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 164–167
April 19, 2021. ©2020 Association for Computational Linguistics

3 Experiments

The HopeEDI dataset consists of 3 languages: English, Tamil and Malayalam. The task can be addressed as a sequence classification problem with 3 class: Hope Speech, Not Hope Speech and Not belonging to the given language. Weighted F1 score is employed as an evaluation metric over the 3 languages separately. Moreover, Tamil and Malayalam consist of text samples in romanized and native scripts. This section describes the experiments conducted for the task and states their outcomes over the validation set.

3.1 Transfer Learning

Customarily, models involving NLP Tasks were trained after random initialization of the network parameters. In Transfer Learning, a neural network is fine-tuned on a particular task after being pre-trained on a general task enabling a given neural network to converge faster and lesser amount of data. Originally, transfer learning has been mainly linked with the fine-tuning of deep learning models trained on the ImageNet dataset(Deng et al., 2009). Recently, the field of NLP has witnessed the emergence of various transfer learning techniques and architectures which considerably improved upon the state-of-the-art on a wide array of NLP tasks. Transfer learning can be employed for applications where there is a lack of availability of sufficient training data. The target dataset should ideally be related to the priorly trained dataset for effective learning. This nature of training is generally attributed as Semi-Supervised training where the network is first trained as a language model on a comprehensive dataset followed by supervised training on a labelled training dataset.

We evaluate such models for the task of hostility-diffusing speech detection, trained over a batch size of 128 over 4 epochs. For English, we experiment with BERT(Devlin et al., 2019) and RoBERTa(Liu et al., 2019).

Model	F1-Score
BERT-base-cased	0.9230
BERT-large-cased	0.9253
RoBERTa-base	0.9313
RoBERTa-large	0.9261

Table 1: Weighted F1 score over English Validation Set

3.2 Paraphrasing is not always Adversarial

Adversarial perturbations attempt to fool models by feeding deceptive input. In general, when a data sample is perturbed, they appear to maintain the same fidelity for humans but manage to get the confuse model prediction. The model mispredicts the target for the perturbed sample as opposed to predicting correctly in the original scenario.

Paraphrasing(Lei et al., 2019) is one such attack that preserves both semantic meaning and syntactically validity as well as transforms text into suitable replacements. However, we apply an earlier variation of sentence-level paraphrasing(Mallinson et al., 2017) as a means for language-targetted data augmentation.

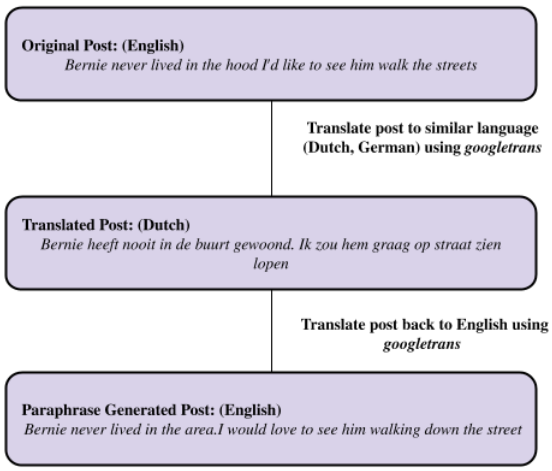

Figure 1: Paraphrase Generation

In this approach, we consider Dutch as the translating language. Both English and Dutch fall under the West Germanic hierarchy of Proto-Germanic languages, with English being Anglo-Frisian and Dutch being Netherlandic-German. Therefore, by choosing a similar language, the generated paraphrases do not lose much semantic meaning but add variability for effective data augmentation. It may noted, this approach cannot be extended to other languages in the task such as Malayalam and Tamil, because translation method (googletrans) employed cannot handle romanized and transliterated text effectively.

Approach	F1-Score
BERT + Paraphrasing Aug.	0.90
RoBERTa + Paraphrasing Aug.	0.90

Table 2: Weighted F1 score over Test Sets using Paraphrasing Data Augmentation

4 Results

Following are described results of the task evaluated over the test set. Our system ranked 1st overall with F1-Score of 0.93 for English, and 4th for Malayalam with 0.78 F-Score.

Model	F1-Score
Baseline (Chakravarthi, 2020)	0.90
roBERTa-base	0.90
BERT-large-cased	0.88

Table 3: Weighted F1 score over English Test Set

Model	F1-Score
Baseline (Chakravarthi, 2020)	0.73
mBERT-cased	0.81
XLM-RobERTa-Large	0.81

Table 4: Weighted F1 score over Malayalam Test Set

5 Conclusion

In this work, we present a simple means of Hope Speech Identification using Pre-trained transformers and Paraphrasing Generation for Data Augmentation. Our future work shall concentrate on interpretability, specifically answering questions like what makes a text an instance of Hope Speech. Moreover, we will attempt to couple more modalities with text, such as audio recording and images or even video clips that collectively promote hope speech.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(0):135–146.

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, M Anand Kumar, John Philip McCrae, Premjith B, Soman KP, and Thomas Mandl. 2020. 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, 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.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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.

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.

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alex Dimakis, Inderjit S. Dhillon, and Michael J Witbrock. 2019. Discrete adversarial attacks and submodular optimization with applications to text classification. In *Proceedings of Machine Learning and Systems*, volume 1, pages 146–165.

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.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

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. 2020. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889. IOS Press.

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.

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.