giniUs@LT-EDI-ACL2022: Aasha: Transformers based Hope-EDI

Basavraj Chinagundi*
bchinagundi_be19@thapar.edu
Thapar Institute of Engineering and Technology, Patiala

Harshul Surana*
harshul19@iiserb.ac.in
Indian Institute of Science Education and Research, Bhopal

Abstract

This paper describes team giniUs' submission to the Hope Speech Detection for Equality, Diversity and Inclusion Shared Task organised by LT-EDI ACL 2022. We have fine-tuned the RoBERTa-large pre-trained model and extracted the last four Decoder layers to build a binary classifier. Our best result on the leaderboard achieves a weighted F1 score of 0.86 and a Macro F1 score of 0.51 for English. We rank fourth in the English task. We have open-sourced our code implementations on GitHub to facilitate easy reproducibility by the scientific community.

1 Introduction

"Hope is a good thing, maybe the best of things, and no good thing ever dies." - Andy Dufresne [Shawshank Redemption]

Hope, as defined by Wikipedia, is an "optimistic state of mind that is based on an expectation of positive outcomes with respect to events and circumstances in one's life or the world at large." Hope is important in life, and research shows that it reduces the feeling of helplessness, helps manage stress and anxiety, cope with adversity, increases happiness and inspires positive action (Chakravarthi, 2020).

The rise of the internet and Social Media has brought the world closer and enabled improved communication and interaction among people. The various forms of interaction include, but are not limited to, online blogs and comments on social media sites like Youtube, Reddit, Facebook, et cetera (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022b; Bharathi et al., 2022; Priyadharshini et al., 2022). The downsides are that Studies have reported that people who excessively use the Internet spend less time interacting face to face, resulting in depression and loneliness (Ybarra et al., 2005). The presence of hate, abuse and discrimination in online interactions is widely studied and documented. While it is essential to highlight and redress these pertinent issues, it is also imperative to highlight the presence of positive online interactions, which can serve as examples of good conduct and etiquette and inspire positive action from the online community (Chakravarthi et al., 2021). The given task involves text classification. Text classification is a classical problem in natural language processing (NLP), which aims to assign pre-defined labels or tags to text, including sentences, queries, paragraphs and documents. It plays an essential role in a multitude of applications such as sentiment analysis, topic labelling, question answering and spam detection.

Historically, rule-based and statistical models have been used to classify texts in the last five decades. The popular techniques include Bag of Words for rule-based and Naive Bayes, Support Vector Machines, and Random Forest for statistical methods. Since the 2010s, text classification has gradually incorporated more deep learning techniques (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). The NLP community has witnessed many innovative architectures like RNNs, LSTMs, and GRUs that push the boundaries of the SOTA. The latest and most impactful is the Transformer, which further gave rise to SOTA models like BERT, RoBERTa, and ALBERT.

2 Task Description

The Hope Speech Detection for Equality, Diversity and Inclusion task (Chakravarthi and Muralidaran, 2021; Hande et al., 2021; Chakravarthi et al., 2022a) aims at identifying Hope Speech. Hope Speech, for the given task, is defined as "YouTube comments / posts that offer support, reassurance, suggestions, inspiration and insight". Hope Speech Detection is an integral component under the overall theme of making Language Technologies more equitable, diverse and inclusive. The task is offered in the following languages - English, Tamil, Span-

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 291 - 295

ish, Kannada, and Malayalam. We participated in the English language task. This is the second edition of the Hope EDI Shared task.

2.1 Dataset

The English language dataset provided by the organizer (Chakravarthi, 2020) consists of Training, Development, and 2 Test sets (Test and New Test). The training set contains 22,740 comments. 20,778, which constitutes 91% of the train set, are examples of non-hope speech, and the remaining 1962 are instances of hope speech. This highlights the heavily imbalanced nature of the dataset and the peculiar challenges it poses as a research question. The Development, or the Validation set consists of 2841 data points. The distribution of hope and non-hope speech is almost the same as the train set (90% and 10% respectively). The test sets contain 2843 and 389 unlabeled instances respectively.

Sample speeches can be found in fig.[2]

3 Setup And Approach

3.1 Experimental Settings

We used the Google Colab's Tesla P100-PCIe-16GB with 8 core CPU and 32GB RAM for training and inference. RoBERTa Decoderizer's max length was set to 22, according to the mean length of sentences after tokenizing. We set the learning rate as 2e-5 and Adam epsilon value as 1e-8 as our Adam Optimizer hyperparameters. We chose an appropriate loss function BCEWithLogitsLoss() for the task. The model was trained for 3 epochs and the best weights were used for the final testing on the Test set.

3.2 Methodology

The text is pre-processed minimally to ensure low information loss in three steps. Firstly, the Unicode characters are removed, after which the domain URLs are removed, followed by the lower casing in the final step. This task aims to identify whether a comment contains hope speech or not and for this, we come up with a Transformer-based approach. The Transformers (Puranik et al., 2021) are designed to take the entire input sentence at once. The primary reason for constructing a Transformer was to enable parallel processing of the words in sentences. This concurrent processing is not possible with LSTMs, RNNs, or GRUs as they take words from the input phrase one at a time. Consequently, in the encoder part of the Transformer, the very first layer has the number of units equal to the number of words in a sentence, and each unit converts that word into an embedding vector corresponding to that word. This allows a better contextual feature extraction of the text, enhancing the ability to determine if the speech is inducing hope. We experiment with prominently known models namely BERT-base-uncased, RoBERTa-base, RoBERTa-large (Liu et al., 2019). We find that RoBERTa-large performs the best when the last four layers of the language model are concatenated for a deeper embedding representation, which is then passed through a pre-classifier and a ReLU activated layer followed by a dropout layer before finally coming across the classification head for the labels that are to be predicted.

4 Results and Discussion

We observe that there is a non-trivial improvement in our model with respect to the RoBERTa-large model. This is because of the novelty of concatenating the vectors of the last 4 layers of the Transformer Decoder. We received this inspiration from the BERT paper (Devlin et al., 2018). We achieve a Macro F1 score of 0.47 and a weighted F1 score of 0.86, and after expanding the feature space by concatenation, we obtain a significant difference in the result. We achieve a Macro F1 score of 0.8 and a weighted F1 score of 0.93. Our best-performing model i.e RoBERTa-large with the last 4 layers concatenated is used for the submission to the leaderboard(lb), and it obtains a Macro F1 score of 0.51 and a weighted F1 score of 0.86. This is competitive with the highest leaderboard results. Results can be found in table[1]. The BERT developers in their paper reported an improved score of 96.1% compared to the baseline of 94.9% to the CoNLL-2003 Named Entity Recognition results. Since the underlying Transformer concept is common to both BERT and RoBERTa, we decided to test if this variation resulted in an improved score. The dataset is observed to be imbalanced and for getting deeper representations of the minority classes, we investigate the impact of the last few layers of the Transformer model. For handling this, we come across the different layers of the model capturing different levels of representations. The layers learn a rich hierarchy of linguistic information i.e. surface-level features in the lower layers, syntactic features in the middle layers, and semantic features in the higher layers. The authors, however, also point out

Model	M_Precision	M_Recall	M_F1-score	W_Precision	W_Recall	W_F1-score
RoBERTa-large(Dev)	0.45	0.5	0.47	0.82	0.9	0.86
RoBERTa-large-last-4(Dev)	0.83	0.77	0.8	0.93	0.94	0.93
giniUs-lb-score	0.51	0.51	0.51	0.86	0.86	0.86
IIITSurat-lb(Highest)	0.56	0.54	0.55	0.87	0.89	0.88

Table 1: Results for Hope Speech Classification

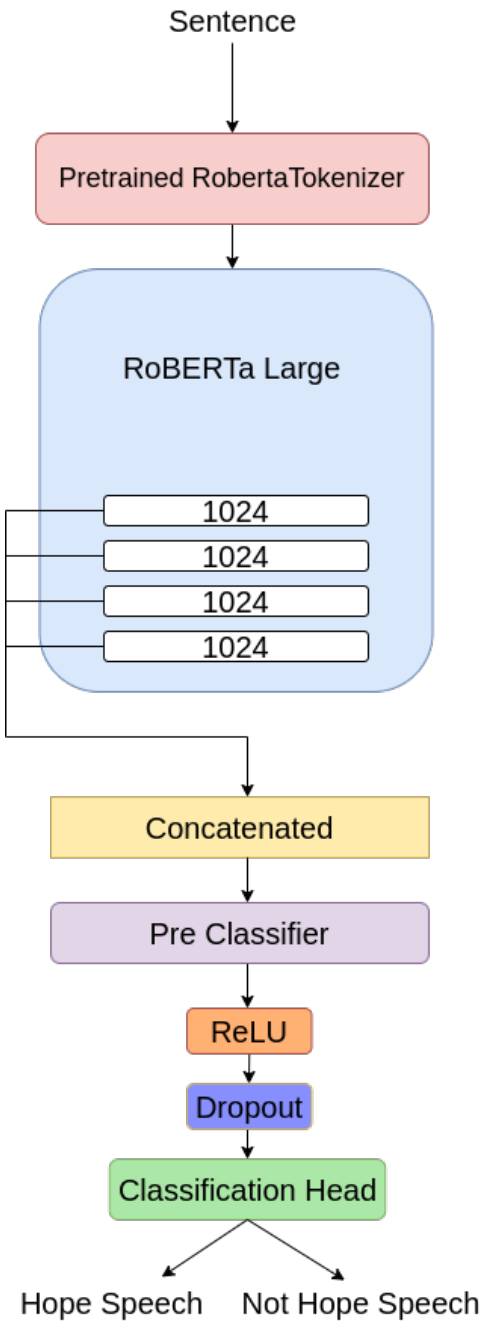

Figure 1: Architecture Diagram

The interviewer is HOT!!!	Non_hope_speech
I am so glad that you made your toy!!!	Hope_speech

Figure 2: Dataset Example

that this technique is not a universal guarantee of improved performance. Instead, it is highly task-specific. In our understanding, unlike many other emotions or sentiments like anger and hate, hope is not a single-dimensional sentiment. Hope constitutes a multiplicity of interpretations that are both personal and complex, thus enabling us to obtain deeper representations of minority class texts and helping us to overcome its downsides.

5 Conclusion

We have presented a novel fine-tuned RoBERTa implementation for the Hope Speech for Equality, Diversity and Inclusion Shared Task ACL 2022. Our best performing model utilises the last four Transformer Decoder layers of the fine-tuned RoBERTa-large model to give a weighted and Macro F1 score of 0.86 and 0.51, respectively. We rank fourth in the leaderboard among all the participants and have released the open-source code for easy and reproducible results.

References

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 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 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*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, John Phillip McCrae, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José Antonio García-Díaz. 2022a. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022b. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language.

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*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2104.09066*.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second 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.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

Michele L Ybarra, Cheryl Alexander, and Kimberly J Mitchell. 2005. Depressive symptomatology, youth internet use, and online interactions: A national survey. *Journal of adolescent health*, 36(1):9–18.