Computer Science
&
Engineering

Mini-Project Report - 6th Semester

Hope Speech Detection for Equality, Diversity, and Inclusion

Anuj Kulkarni 19BCS012,
M.Mallikarjun 19BCS064,
Mahesh Parihar 19BCS066,
Niranjan Meghwal 19BCS078

supervised by:
Dr. Sunil Saumya

Acknowledgement:

We would like to express our profound gratitude to Dr. Sunil Saumya (Assistant Professor, Computer Science and Engineering department) for his contributions to the completion of our project titled Hope Speech Detection for Equality, Diversity, and Inclusion. We would also like to express our special thanks to Mr. Shankar Biradar (PhD Scholar, Computer Science and Engineering department) for his time and efforts provided throughout the semester. Their useful advice and suggestions were really helpful to us during the project's completion. In this aspect, we are eternally grateful to them. This project helped us in doing a lot of research and we came to know about so many new things. We are really thankful to them.

1 Introduction

With the expansion of the Internet, there has been substantial growth all over the world in the number of marginalised people looking for support online. Recently, due to the lockdowns enforced as a consequence of the COVID-19 pandemic, people have started to look at online forums as an emotional outlet when they go through a tough time.Young people who experience online hate are more likely to experience anxiety and depression, and targets of online hate may suffer harassment.It interferes with individual's ability to communicate with others and to empathise. Because it often relies on stereotypes and scapegoating, it negatively impacts our ability to address the root causes of social problems.

Social Media has inherently changed the way people interact and carry on with their everyday lives as people using the internet. Due to the vast amount of data being available on social media applications such as YouTube, Facebook, and Twitter it has resulted in people stating their opinions in the form of comments that could imply hate or negative sentiment towards an individual or a community Therefore hate speech can also harm communities, even when it targets individuals. This results in people feeling hostile about certain posts and thus feeling very hurt.

Being a free platform, social media runs on user-generated content. With people from multifarious backgrounds present, it creates a rich social structure and has become an exceptional source of information. It has laid its roots so deeply into the lives of people that they count on it for their every need. Regardless,this tends to mislead people in search of credible information. Certain individuals or ethnic groups also fall prey to people utilising these platforms to foster destructive or harmful behaviour.

1.1 Motivation

Over the past few years, systems have been developed to control online content and eliminate abusive, offensive or hate speech content. However, people in power sometimes misuse this form of censorship to obstruct the democratic right of freedom of speech. This has also led to controlling user expression instead of improving user experience.Such harmful content could spread, stimulate, and vindicate hatred, outrage, and prejudice against the targeted users. Removing such comments was never an option as it suppresses the freedom of speech of the user and it is highly unlikely to stop the person from posting more. Therefore, it is imperative that research should take a positive reinforcement approach towards online content that is encouraging, positive and supportive. Until now, most studies have focused on solving this problem of negativity in the English language, though the problem is much more than just harmful content.We should turn our work towards spreading positivity instead of curbing an individual's freedom of speech by removing negative comments. Therefore, we turn our research focus towards hope speech.

To cherish a desire with anticipation is one of the definitions of Hope. Hope is considered significant for the well-being, recuperation and restoration of human life by health professionals. Hope can be defined as an optimistic state of mind that depends on a desire for positive results regarding the occasions and conditions of one's life or the world at large, and it is also present and future-oriented.Hope can also come from inspirational talk about how people face difficult situations and survive them. Hope speech engenders optimism and resilience that positively influences many aspects of life. We define hope speech for our problem as "YouTube comments/posts that offer support, reassurance, suggestions, inspiration and insight".

2 Related Work

With the fast growing world of social media and it's users there is a scope for learning behaviours and incentivizing appropriate ones. Since the gap between the message sender and receiver apparently has decreased people want to connect to more people without realising they are talking to mere screens, hence sometimes this leads to certain situations where one might cross a line without even bothering about how the other might react since there online presence is negligible. This provides an opportunity to analyze and collect related data so as to bring more awareness on human interaction in this global village. Many attempts have been made to go through social media data by crawling, for example *marrese et al. (2017)* [1] did the same for Youtube comments for opinion mining looking at video reviews as evolved form of written product reviews, *Severyn, Aliaksei et al. ACL (2014)* [2] also did opinion mining on Youtube comments using tree kernel technique to extract features, *Romim et al.(2021)* [3] generated a dataset in Bengali language which consists of 30,000 user comments collected from YouTube and Facebook comment section and classified into seven categories: sports, entertainment, religion, politics, crime, celebrity and TikTok and meme. *Shriphani Palakodety et al.(2019)* [4] analyzed evolving international crisis during the tension between India and Pakistan after the Pulwama terror attack via a substantial corpus constructed using comments on YouTube videos. The research in the area of Hope speech itself is quite extensive, while reading through literature we found many examples that had previously worked on Hope speech, for example *Hossain et al.(2021)* worked on finding out whether or not a social media post/comment contains hope speech by using various ML and DL techniques, *Junaida et al.(2021)* [5] [6] did a similar analysis on a multilingual dataset and used DL techniques to find word embeddings which were then used with RNN(s). Indic Languages are often refered to as resource scarce, meaning the amount of research done on them or the availability of dataset poses a problem in further research in the same field, we found some interesting papers which were done on Indic languages, *chakravarthi et al. (2020)* [7] came up with code-mixed datasets on resource scarce indic languages like Tamil, Malayalam for sentiment analysis, similarly *kamble et al. (2018)* [8] reported their advancements in the area of hate speech recognition for English-Hindi code-mixed tweets. Some papers on hate speech detection that we came across are *Nobata et al. (2017)* [9] have worked on Abusive Language Detection whereas, *schmidt et al.(2017)* [10] did a survey on hate speech detection using NLP.

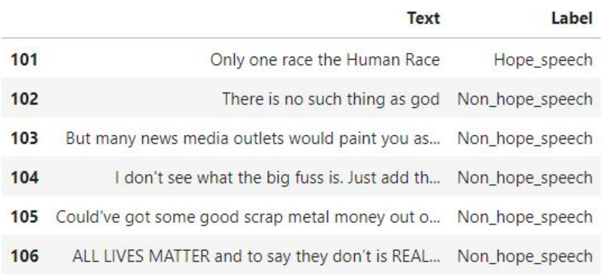
3 Dataset Description

We used the dataset provided in the competition Hope Speech Detection for Equality, Diversity, and Inclusion -ACL 2022 organized by DravidianLangTech-2021 [1]

which was the first workshop on Speech and Language Technologies for Dravidian Languages. Our dataset consists of:

- Hope_ENG_train - used as 'train' data
- Hope_ENG_dev - used as 'test' data

our dataset comprises of two columns 'Text' which contains the comment and 'Label' which stores corresponding label.



	Text	Label
101	Only one race the Human Race	Hope_speech
102	There is no such thing as god	Non_hope_speech
103	But many news media outlets would paint you as...	Non_hope_speech
104	I don't see what the big fuss is. Just add th...	Non_hope_speech
105	Could've got some good scrap metal money out o...	Non_hope_speech
106	ALL LIVES MATTER and to say they don't is REAL...	Non_hope_speech

Figure 1: randomly picked 6 sentences from test dataset

[1] HopeEDI-ACL 2022

Dataset discription

	Test data	Train data
Total no. of sentences	22,740	2840
Length of longest sentence	1025	926
Hope speech	1962	272
Non hope speech	20778	2568

4 Methodology

We followed the below described workflow for our project, for data collection as stated above we found our dataset from Hope Speech Detection for Equality, Diversity, and Inclusion. The dataset was pre-divided into train and test data respectively. Now after data collection comes an important step of cleaning and pre-processing data.

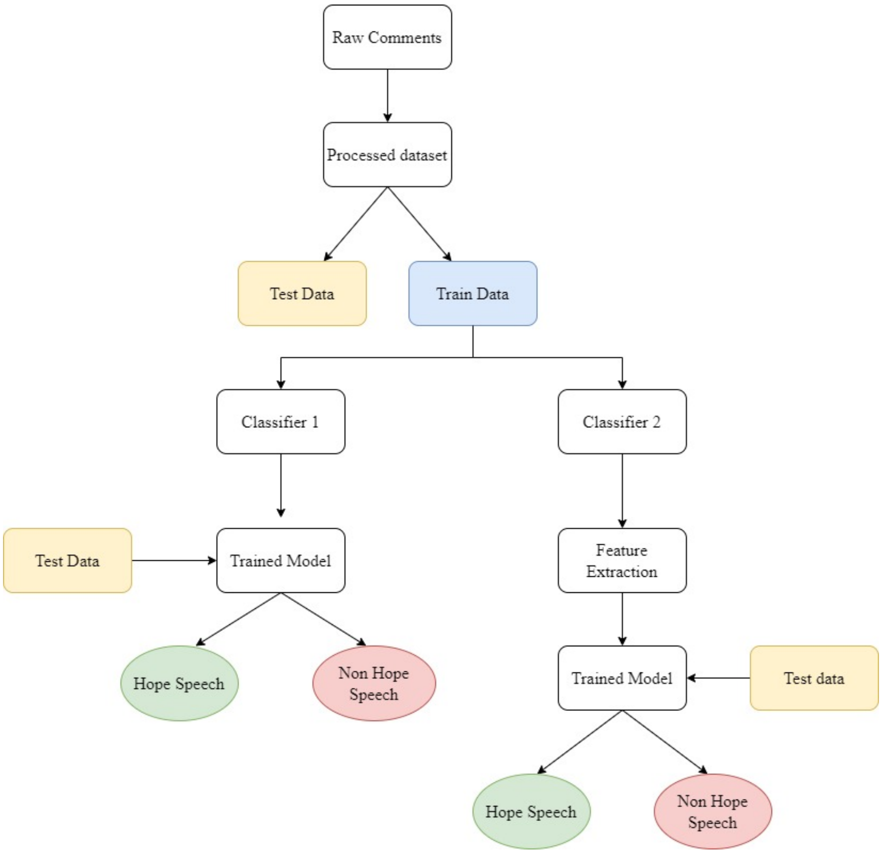


Figure 2: Project Workflow

4.1 Data cleaning and pre-processing

Natural Language Processing (NLP) is a branch of Data Science which deals with Text data. Apart from numerical data, Text data is available to a great extent which is used to analyze and solve problems. But before using the data for analysis or prediction, processing the data is important.Text processing is a method used under the NLP to clean the text and prepare it for the model building. It is versatile and contains noise in various forms like emotions, punctuations, and text written in numerical or special character forms. We have to deal with these main problems because machines will not understand they ask only for numbers. We have cleaned data by converting all words into lowercase, the words themselves store information their variations regarding casing would only increase the number of duplicates. We have also removed HTML syntax if present, such markdown methods might be helpful in presenting the content in an approachable manner but do not add any significance to the overall sentiment itself. If any kind of URL(s) are present then we have removed them also, Our current objective is to figure out techniques to understand and analyze the over all sentiment surrounding the text data itself. Now the we have used lemmatization technique to convert words to their respective root forms(lemma), this would help us to cut down any chances of duplication that might arise due to the different forms or tense related variations of a single word. Now we convert this preprocessed data from group of words to group of tokens, we are able to do so with the help of Tokenizer from Tensorflow keras. Tokenizer helps us to vectorize text corpus by turning each sequence of tokens into a sequence of integers. Now this data is ready to be fed into classifiers.

4.2 Classifiers

- Classifier 1 (BiLSTM):Bidirectional long-short term memory(bi-lstm) is the process of making any neural network o have the sequence information in both directions backwards (future to past) or forward(past to future).

  In bidirectional, our input flows in two directions, making a bi-lstm different from the regular LSTM. With the regular LSTM, we can make input flow in one direction, either backwards or forward. However, in bi-directional, we can make the input flow in both directions to preserve the future and the past information.

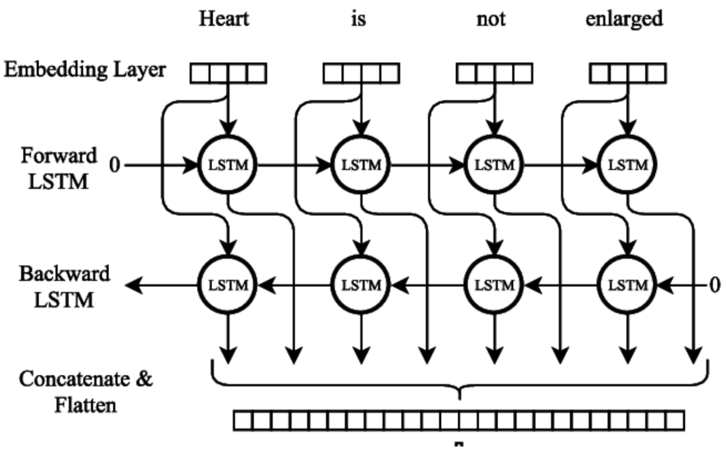


Figure 3: BiLSTM (Ref : *Bakewell et al. (2016) [11]*)

We have cut short the maximum accepted length of sentence to 200. The model was then run for 60 epochs. We have added the following layers to build up our BiLSTM model:

Model: "sequential"

Layer (type)	Output shape
embedding (Embedding)	(None, 200, 128)
bidirectional (Bidirectional)	(None, 128)
dropout (Dropout)	(None, 128)
dense (Dense)	(None, 1)

- Classifer 2 (TF-IDF vectors + SVM):

  Term Frequency: In document d, the frequency represents the number of instances of a given word t. Therefore, we can see that it becomes more relevant when a word appears in the text, which is rational. The weight of a term that occurs in a document is simply proportional to the term frequency.

  $$tf(t,d) = \text{count of t in d / number of words in d}$$

  Document Frequency: It is very similar to TF, the only difference is that in document d, TF is the frequency counter for a term t, while df is the number of occurrences in the document set N of the term t. In other words, the number of papers in which the word is present is DF.

$df(t,d)$ = occurrence of t in documents

Inverse Document Frequency: It tests how relevant the word is. The key aim of the search is to locate the appropriate records that fit the demand. At first, find the document frequency of a term t by counting the number of documents containing the term:

$$df(t) = N(t)$$

where, df(t) = Document Frequency of a term t

N(t) = Number of documents containing the term t

Term Frequency is the number of instances of a term in a single document only: although the frequency of the document is the number of documents in which the term appears, it depends on the entire corpus. Now let's look at the definition of the frequency of the inverse paper. The IDF of the word is the number of documents in the corpus separated by the frequency of the text.

$$idf(t) = N/df(t) = N/N(t)$$

TF-IDF: TF-IDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining and user modeling. The formula that is used to compute the tf-idf for a term t of a document d in a document set is

$$tf - idf(t,d) = tf(t,d) * idf(t)$$

,

and the idf is computed as

$$idf(t) = log[n/df(t)] + 1$$

,

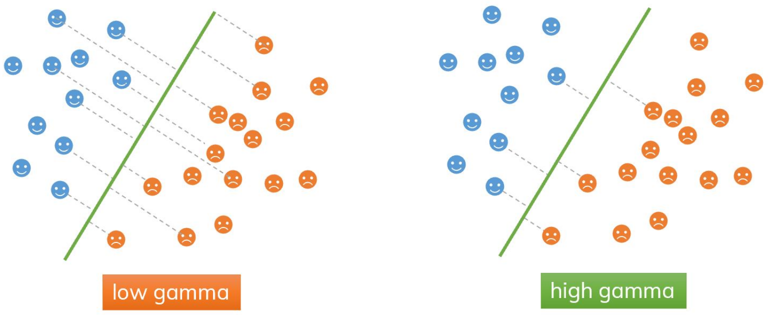
where n is the total number of documents in the document set and df(t) is the document frequency of t.

SVM: Support Vector Machine(SVM) *Khanh Nguyen et al. (2015)* [12] is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features.

Parameters used are:-

Kernel:- SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. The popular kernels are Polynomial Kernel, Gaussian Kernel, Radial Basis Function (RBF), Laplace RBF Kernel, Sigmoid Kernel, Anove RBF Kernel. in our model we will use linear kernel.

Gamma:- The gamma parameter defines how far the influence of a single training example reaches. This means that high Gamma will consider only points close to the plausible hyperplane and low Gamma will consider points at greater distance.



decreasing the Gamma will result that finding the correct hyperplane will consider points at greater distances so more and more points will be used (green lines indicates which points were considered when finding the optimal hyperplane).

Margin:- The last parameter is the margin. We've already talked about margin, higher margin results better model, so better classification (or prediction). The margin should be always maximized.

5 Results

After applying the above discussed methods and techniques we have found the following results :

Scores	Classifier 1 (BiLSTM)	Classifier 2 (TF-IDF vectors + SVM)
Accuracy	86.87 %	91 %
Precision	85.09 %	89 %
f1 score	85.91 %	89 %

It is observed that Classifier 2(TF-IDF vectors + SVM) gives better results than Classifier 1(BiLSTM).

6 Limitations and future scope

Our dataset shows a great imbalance between Non Hope Speech and Hope Speech examples, this creates a problem while accurately assessing the model's accuracy, there is a need of a more balanced dataset preferably with higher total number of examples. The recent pandemic has given rise to the fast growing online community with increase in online presence, more people more opinions, this can be seen as an opportunity to use methods such as web scrapping and comments crawling to create a robust dataset which might be a step towards novelty in this field of study. While discussing about data preprocessing we highlighted the need of tokenizer, tokenizer also removes punctuation and emoticons, this helps in segregating word tokens from other unnecessary information or noise but we know that emoticons and punctuation do add value to the overall sentiment of the statement. URL(s) if present might become a gateway to understand the actual intent of the comment/review writer but were removed completely during our project such neglect can give rise to an incorrect classification. The recent advancements in the field of 'Transformer' architecture has given rise to BERT model which can provide word embeddings while considering punctuation, emoticons, emojis etc. these can be used to analyze the effectiveness of our classification model. We need to find a way to include URL(s) etc. to the analysis of overall sentiment to provide better classification.

7 Conclusion

In the era of fast growing internet, public opinion on a rapidly evolving global issue can exhibit similar fast-changing behavior, much of which is visible to a very large fraction of internet users, it is necessary to encourage positivity such as in the form of hope speech in online forums to induce compassion and acceptable social behaviour. We aim to encourage positive content in online social media for equality, diversity and inclusion. This project presents some methodologies that can detect hope in social media comments.We have used two classifiers that are bilstm and tf idf vectors +svm. We also conclude that in this case the tf idf vectors + svm model have given better results than bilstm. In the future, we plan to extend the study by trying out different and more advanced techniques with further fine-grained classification and content analysis to improve our results.

References

[1] E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo, "Mining fine-grained opinions on closed captions of youtube videos with an attention-rnn," *arXiv preprint arXiv:1708.02420*, 2017.

[2] A. Severyn, O. Uryupina, B. Plank, A. Moschitti, and K. Filippova, "Opinion mining on youtube," 2014.

[3] N. Romim, M. Ahmed, H. Talukder, S. Islam *et al.*, "Hate speech detection in the bengali language: A dataset and its baseline evaluation," in *Proceedings of International Joint Conference on Advances in Computational Intelligence.* Springer, 2021, pp. 457–468.

[4] S. Palakodety, A. R. KhudaBukhsh, and J. G. Carbonell, "Hope speech detection: A computational analysis of the voice of peace," *arXiv preprint arXiv:1909.12940*, 2019.

[5] M. Junaida and A. Ajees, "Ku_nlp@ lt-edi-eacl2021: a multilingual hope speech detection for equality, diversity, and inclusion using context aware embeddings," in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, 2021, pp. 79–85.

[6] E. Hossain, O. Sharif, and M. M. Hoque, "Nlp-cuet@ lt-edi-eacl2021: multilingual code-mixed hope speech detection using cross-lingual representation learner," *arXiv preprint arXiv:2103.00464*, 2021.

[7] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, and J. P. McCrae, "A sentiment analysis dataset for code-mixed malayalam-english," *arXiv preprint arXiv:2006.00210*, 2020.

[8] S. Kamble and A. Joshi, "Hate speech detection from code-mixed hindi-english tweets using deep learning models," *arXiv preprint arXiv:1811.05145*, 2018.

[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.

[10] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain.* Association for Computational Linguistics, 2019, pp. 1–10.

[11] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," *arXiv preprint arXiv:1609.08409*, 2016.

[12] K. Nguyen, T. Le, V. Lai, D. Nguyen, D. Tran, and W. Ma, "Least square support vector machine for large-scale dataset," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.