Proceedings of the Sixth International Conference on Intelligent Computing and Control Systems (ICICCS 2022)
IEEE Xplore Part Number: CFP22K74-ART; ISBN: 978-1-6654-1035-9

Hate Speech Detection Using Machine Learning In Bengali Languages

2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1035-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCS53718.2022.9788344

1st Mominul Islam
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
mominul15-11992@diu.edu.bd

2nd Md Sanjid Hossain
dept. Of CSE
Daffodil International University
Dhaka, Bangladesh
sanjid15-11888@diu.edu.bd

3rd Nasrin Akhter
Daffodil International University
Dhaka, Bangladesh
nasrin.cse@diu.edu.bd

***Abstract*—Hate speech is a common problem in the current time of social media and the internet as it is very easy to be in touch with everything through the internet and social media. Hate speech detection research is not very rare but in terms of Bengali language there are very few works related to hate speech in Bengali language. The proposed research experiment has developed a machine learning based project to detect hate speech from Bengali language data or comments, posts in social media that are in Bengali language. This research work has used 3006 pure Bengali data from social media pages (such as Facebook, YouTube) groups, comment sections of news portals. Further, this research work has categorized them in 0 for non-Hate-Speech and 1 for Hate-Speech to classify the data between non-abusive and abusive data. This research work has used several algorithms to find the best possible result in order to determine whether the sentence is abusive or non-abusive such as Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, K Nearest Neighbor Classifier. From these algorithms, the best result for detecting non-abusive data is the Random Forest [RF] algorithm, which is 67%.**

Keywords—Term frequency-inverse document frequency [TF-IDF], Count Vectorizer, Machine Learning Algorithms, cleaning data, annotation data, Abusive Speech

I. INTRODUCTION

The position of Bengali Language is the 6th among the most spoken languages all over the world [11], where it is spoken by almost 260 million people. Which creates more potential and possibilities in terms of language and speech detection. Bengali language is the national language of Bangladesh and also spoken in India, Bengali language is the second most-spoken language [12]. Since the declaration of UNESCO, we have observed 21st February as International Mother Language Day. Almost all the people of the world have to respect their own language. About 47 million people are using Facebook in Bangladesh [13] Which is about 28 percent of the total population of the country. Of these users, 61 percent are men and 31 percent are women. Moreover, many of them use Twitter, Instagram, YouTube, etc. compared to previous years, Internet packs are much cheaper now in Bangladesh. Since the internet has reached a massive part of our society, the use of social media has also increased which has also increased the line of communication and freedom of speech has reached a different level. Now with a greater number of increased users, there has also been an occurrence of many problems over the internet. Since access to the internet is so easy people can easily express themselves on social media and similar platforms and with that many opportunities there are also some forms of negativity. With this much freedom there are a large number of people that actually abuse power and opportunity the wrong way. Some people use the social media coverage to spread rumors and defamation information on social media, while few groups directly use abusive language against certain groups or peoples and there are also some people that use the internet and social media to violate others right by using hateful language, inappropriate and sexual comments. On top of that, all the above-described social media are public. Now according to some sources [9] there are a total 46 million Facebook users in Bangladesh which is more than the number of users in Pakistan (45 million). This is only the number of users that use Facebook currently. Apart from that, there is also YouTube which is not a social media but it has a great influence on people since it is also a public platform that can be used for expression of free speech. Using only these following sources for data is great but there is also a new door of possibility of data collection which is comment section YouTube and news portals [10] According to a report by the Daily Star, in 2018 there were over 29 million YouTube viewers. While The number has now increased more significantly since covid-19, this number is still increasing on a daily basis.

Hence, we decided to work on Bengali language, after analyzing all the possibilities and current studies it shows great potential on Language detection, since work on Bengali language specific are very rare and it is a growing and one of the most used languages and most of the work related to Bengali language was on mixed language detection, our primary goal was to work on complete and pure Bengali language. Our secondary goal, was to ensure the quality of the data as only Bengali language specific data was rare and so were the sources.

In our research we have developed and worked on multiple machine learning (ML) based algorithms to detect abusive language and hate speech from Bengali language by using several classifiers on the cleaned and annotated data from

social media and public platforms. We have used many social media platforms to collect the data. And since Facebook is the biggest and most used social media platform, we have collected most of our data from Facebook and we have also collected data from YouTube since YouTube is also one of the most used public platforms where people also express themselves. On top of that we have also taken data from news portals. Since we had targeted to use 7000 data, we have collected almost 40000 data from both platforms at the beginning. The data we collected was mixed with various unnecessary data elements and attributes such as emojis, mix language, images, URL and some garbage values. So, to ensure the purest data, we collected more than 10000. Purity was a necessity in our research as the fact that there was not much research and resources on pure Bengali language. Therefore, we had to manually do most of the work to get the best result. So, we collected more than 40000 data and then we started the annotation, data cleaning and other processes.We have described and demonstrated a small portion of our dataset in this paper while we have stored and posted [16] our dataset publicly.

II. Literature Review

In the paper [1] the author conducted Detect Abusive Bengali language text. In this paper, the dataset is three-class, positive, neutral, negative. The author gained the best performance in the RNN algorithm to compare other machine learning and deep learning algorithms. This paper improved text classification to detect which text is a personal attack, politically violated, antifeminism abuse. [2] From an article to detect hate speech worldwide, we have discovered that hatred or some term of hate speech can be determined by characterizing them indifferent sets of group while some of the hate speech are common but each group of hate speech can be determined by using different sets of high frequency stereotypical words.[3] From an Italian research on hate speech detection on Facebook, where author have demonstrated a method to detect hate speech on native language, where author have used SVM(support vector Machines) and LSTM(Long Short Term Memory) classifier for Italian language on the basis of multiple learning algorithm. [4] From a research detecting hate speech on social media, author had set up lexical baselines for this task by applying supervised classification method employing as of late released dataset explaining their reason. As features, their framework employs character n-grams, word n-grams and word skip-grams. [5] From a research on Bengali language Hate Speech detection on social media, where author uses many ML algorithms and a deep neural network model data from single social media(Facebook) to determine a suitable and compare the algorithm performance.1D convolutional layers were used for Extraction and encoding of local feature from comments on Facebook and Long Short Term Memory ,attention mechanism and GRU based encoders were used to predict the hate speech categories.[6] From another research about approach to detect abusive Bengali language text on similar datasets, author have proposed an older approach as the resource in Bengali language is limited. In that research, to detect text, authors have proposed a root level algorithm and to enhance the result they also used unigram string features.[7] A similar research in similar data set to detect hate speech in Bengali language has also developed another approach, as authors have developed Machine Learning algorithm based model, GRU and Deep neural network model to classify data in their research. In their research, they developed their method-ology of dataset collection, annotation and divided their dataset in 6 class and generate topmost uni-gram,tri-gram and bi-gram feature for the classes.[8]From a brief research in similar work cyberbullying identification and tackling using NLP, in attempt to precise the result, the author have developed a software tool for Automatic Detection of Cyberbullying using ML based methodology with 6 class and 5 perceptible feature engineering.The author have used word sense disambiguation with WordNet-aided semantic expansion approach to train dataset through data augmentation and proposed the approach of using NegEx and POS tagging for negation handling.

III. Methodology

A. Dataset collection

To collect data, we used Instant Data scraper software to directly collect data and convert them in CSV file format. Instant scraper works as an extension for Google Chrome. To use this software, the Search address or link needs to be altered.In the web address of targeted social media , in place of ”WWW” a keyword(“mbasic”) is used. Once this process is done the instant scraper can automatically collect all the comments from the page and store them in CSV file format, which can directly be downloaded or shared. We collected data from social media groups, local Bengali news channels pages, political news posts, roasting videos, funny content, and interviews posts. While collecting comments was simple but unlike bot, we had to manually change pages and copy or extract data each time but collecting reply data from them was a bit tough since those were not detected by the software Support Vector Machine and had to be collected manually some time.We have collected about 40,000 data from Facebook. From those 40,000 data, there was a massive amount of data that was neutral. Furthermore, there were also few comments with URL, emojis and images that were unusable to our required data. We manually selected and deleted all the emojis, URLs and images that were unusable. After reducing all the unusable and unnecessary data, we have managed to purify and reduce total data to 7000 data. Which we have achieved through a data cleaning process explained in later steps.

B. Dataset annotation

After data collection, the first most complex task of our research paper was data annotation. After manually deleting Some repeated, Bengali-English mixed, English data, HTML(hyper-text markup language) markups, links, image titles, special characters etc.

TABLE I. DATA LABELING

Data	Label
দুর চুতমারানি,.... এহনো সময় আছে, সাংবাদিক আর নায়িকা দুইডাই ভালা হ	1
আজ রাতে তাকে পুঁটকি মারবে এটা নিয়ে নিউজ করেন	1
এক টিকেটে দুই ছবি গরম মশল্লা	0
এটা কোনো শিরোনাম হলো	0
সালার চামচামির একটা সীমা থাকা উচিত	1
তাতে আমার বাল ছেড়া গেলো	1

At the beginning we started by collecting 40000 data from various sources using scrap. Then we went through the data cleaning process. In the data cleaning process, we examined and erased those data, after that we were left with 7000 data. Which was Pure usable data.

After the data cleaning process, we were left with 7000 data. From which we started our data annotation process carefully with precision. To get precise results from annotated data we complete our labeling process manually. We assign our labeling data into 0(Not Hate speech) and 1(Hate speech). For neutral data we used 0 (Not Hate speech). We used 1 (Hate speech) for labeling data into abusive and hate speech. We assigned all kinds of hate speech and abusive language as 1 (Hate speech). We completed the whole process manually as it was the most crucial part of our research to get the most efficient result, as the fact that our result is dependent on our data annotation process.

After labeling the data, for better results to avoid complexity we took 3006 data that we assigned and categorized as hate speech or abusive language and neutral data or non-abusive.

Here we have presented a small part of our data set in the above table, we can see that there are 4 data. 2 of the first 4 data are labeled as 1 and remaining two are assigned as 0. We labeled them as 1 by determining and confirming some word that is considered as hateful or abusive language in Bangla. Although 2 of the words have different meanings in Bengali language but they are both abusive in different ways on the other hand the remaining two data are categorized as 0 because we can see there was not a single word that had any abusive or hateful meaning in them.

C. Data Preprocessing

As a step of preprocessing, we collected comments which contain only Bengali language using Instant Data scraper software. But, as discussed before, there was a massive amount of unusable data. So, in preprocessing, all types of whitespaces, emoticons and digits from the dataset were removed. Furthermore, to make more enhancement in data we approached a data cleaning process.

In the preprocessing phase, we have also gone through calculation and statistical analysis on our 3006 data we selected for the process where have manually calculated our data and then created a statistical result to show the percentage of abusive and non-abusive data which we can see in fig1(statistical analysis of Hate Speech & non-Hate-Speech). We can see that from 3006 data there were 56.25%, which means 1691 data that was assigned as 0(neutral or non-abusive or non-Hate-Speech) on the other hand there were 43.75% or 1315 data were assigned as 1(abusive and hate-Speech). Which also established the fact that most of the data was neutral or non-abusive.

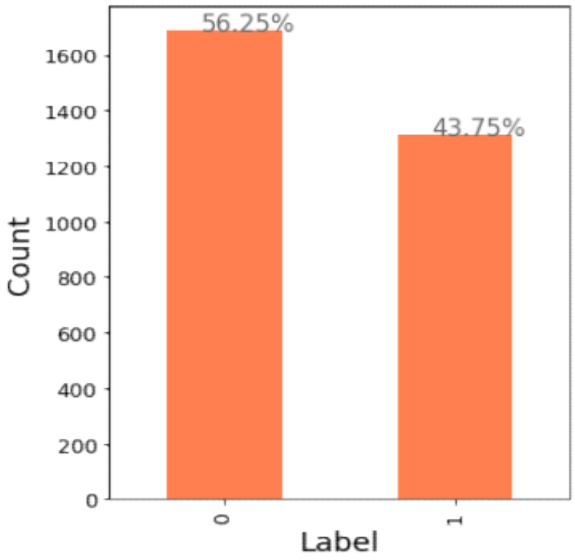

Fig. 1. statistical analysis of Hate-Speech & non-Hate-Speech

D. Data Cleaning

Collected data usually contains all types of elements of texts. For example, all the data contains some form of punctuation like (“। ”, ”,” , ”!”, ”;”) even in Bengali language data, on top of that there is also much data with numeric values(0-9) or numbers. Which is neither necessary nor the usable part of the research. So they were considered as garbage values for our research and possible complication for the research. As a result of that concern, to deal with that issue the Data Cleaning process was demonstrated in table 2.2. Here, cleaning is visible as seen from the comparison, from the first example of the table, before cleaning, there was “???” signs in between the word “কেন” and “আইন”. which was removed through a data cleaning process as demonstrated in “After cleaning” there is no more there is no ”???” signs in between “কেন” and “আইন” . Which is a small part of our data cleaning process

TABLE II. BEFORE & AFTER CLEANING PROCESS

Before Cleaning	After cleaning
এক দেশে একই অপরাধের জন্য দুই আইন কেন?? আইন সবার জন্য সমান হওয়া উচিৎ। এই ঘটনার তীব্র নিন্দা জানাচ্ছি।	এক দেশে একই অপরাধের জন্য দুই আইন কেন আইন সবার জন্য সমান হওয়া উচিৎ এই ঘটনার তীব্র নিন্দা জানাচ্ছি
দুর চুতমারানি..... এহনো সময় আছে, সাংবাদিক আর নায়িকা দুইডাই ভালা হ	দুর চুতমারানি এহনো সময় আছে সাংবাদিক আর নায়িকা দুইডাই ভালা হ
তদন্তের সম্ভাব্য ফলাফল: ১. জামায়াত শিবির সরাসরি জড়িত। ২. বি এন পি উস্কানিদাতা। ৩. মাদ্রাসার ছাত্র শিক্ষক জড়িত।	তদন্তের সম্ভাব্য ফলাফল জামায়াত শিবির সরাসরি জড়িত বি এন পি উস্কানিদাতা মাদ্রাসার ছাত্র শিক্ষক জড়িত

E. Data Transformation

Text after removing punctuations from text, we extracted our text data with count-vectorizer and Term frequency-inverse document frequency vectorizer by these two methods.
These two methods converted machine-readable data from our text data.

CountVectorizer is a python based scikit-learn library. Scikit-learn is an independent python library for machine learning. It supports Python numerical and scientific libraries such as NumPy and SciPy, as well as numerous techniques such as support vector machine, random forests, and k-neighbors.[15] With this tool, we can transform data(text) into vector form based on counting each word of the entire text. With this tool we can convert multiple words from multiple data into vectors for more efficient text/data analysis.[13]

From our text data, we have used the Countvectorizer to generate a matrix where each unique word is represented by a column of the matrix and a row of each text sample matrix from the document. To be more brief, while certain encoding declares a whole sentence as a number, in the case of the count vectorizer, this method declares each word of the sentence into a number instead of a whole sentence. by declearning or assigning values for each word and then from those values, the matrix is generated

TF-IDF (term frequency-inverse document frequency) is an information retrieval and document search method that is used for word scoring in machine learning algorithms for Natural Language Process and automated text analysis. Tf-IDF is a measurable degree that assesses how important a word is to a data in a collection of data.[14]

Term frequency-inverse document frequency (TF-IDF)=Term Frequency (TF) * Inverse Document Frequency (IDF)

Where:
Term Frequency TF= number of repetition words in sentence / Total word in sentence

Inverse Document Frequency (IDF) = Total number of sentences / No of sentences certain in the word.

F. Implementation

To test the performance of the algorithm we have proposed, we first used supervised learning to supervise the data that we have collected from various sources and then we labeled our data into 0 and 1 to determine non-abusive and abusive. We have used two methods, TF-IDF and Count Vectorizer to convert our text data. To determine the accuracy, we have used 5 classifiers on both methods: Logistic Regression, Naive Bayes, RandomForest, SVM (support vector machine), and KNeighbours Classifiers. Pseudocode of a random forest algorithm can be splitted into two stages. One being the Random forest creation pseudocode while the split is pseudocode for performance prediction for the created random forest classifier. For the first split of pseudocode, “k” has to be selected from total features of “m”, where “k<<m”. node “d” has to be calculated for the best split point from “k” features with the use of best split,nodes have to be splitted into daughter nodes. All these processes have to be running till the “l” number of nodes are reached. By repeating all the stages from “k” selection to “l” searching, the forest will be built to create ”n” number of trees.

From Table:3 & Table:4, we can see the precision, recall, f1-score and accuracy for the 5 algorithms that we have used. In table 3, we can see the result of TF-IDF for 5 classifiers. As for table 4, here we can observe the result of Count Vectorizer also for the same 5 classifiers.

First, we used confusion matrix to determine the accuracy between algorithm and labeling where we use four matrices as

TP (true positive): the number of data that classifier correctly predicted as abusive and match label

TN (true negative): the number of data that classifier predicted as abusive and did not match the label

FP (false positive): the number of data that classifier predicted as non-abusive and did not match the label

FN (false negative): the number of data that classifier predicted as negative and match the label

Using the following four terms we have determined the accuracy of our algorithm in both methods. We have

determined precision, F-measure(f1-score), recall and accuracy.

Precision is used in both methods to determine the correctly labeled amount of data in the test set; it determines the precision of the algorithm by dividing
TP with TP and FP.

Precision= True Positive/(True Positive +False Positive)

Recall, is also used in both methods, where it is the total amount of data in the test set that is correctly labeled by classifiers and that are actually labeled for specific class, it is determined as:

Recall= True Positive /(True Positive + False Positive)

F1-measure, also known as the weighted harmonic mean of precision and recall of a given class.

F1-measure=(2*precision*Recall)/(precision+Recall)

Accuracy is the most vital part of the algorithm as it is the total percentage of the data that the classifier has correctly labeled. It is calculated as;

Accuracy=(True Positive + True Negative)/(True Positive + True Negative +False Positive + False Negative)

TABLE III. RESULTS FOR THE ALGORITHM IN TF-IDF

Algorithms	**Scores**			
	precision	***recall***	***F1-score***	***accuracy***
LogisticRegression	***0.64***	***0.64***	***0.63***	***0.64***
NaiveBayes	***0.66***	***0.66***	***0.64***	***0.65***
RandomForest	***0.68***	***0.68***	***0.67***	***0.67***
Support Vector Machine	***0.64***	***0.62***	***0.61***	***0.64***
KNeighbours Classifieer	***0.52***	***0.53***	***0.52***	***0.53***

TABLE IV. RESULTS FOR THE ALGORITHM IN COUNTVECTORIZER

Algorithms	**Scores**			
	precision	***recall***	***F1-score***	***accuracy***
Logistic Regression	***0.63***	***0.64***	***0.63***	***0.63***
NaiveBayes	***0.66***	***0.64***	***0.62***	***0.63***
RandomForest	***0.63***	***0.63***	***0.62***	***0.63***
Support Vector Machine	***0.64***	***0.64***	***0.63***	***0.64***
KNeighboursClassifieer	***0.56***	***0.58***	***0.55***	***0.57***

The data set we have used for this experiment is 3006 sets of data. We have used the whole data set of 3006 in this experiment to determine the result of precision, recall, F-measure and accuracy. As we have used TF-IDF(Term frequency-inverse document frequency) and Count Vectorizer we received two sets of results as well. With TF-IDF(Term frequency-inverse document frequency) the best accuracy we have achieved is with the RandomForest algorithm, where the accuracy was 0.67 or 67% which we can see from table:3. As for the CountVectorizer, from table 4, the best result of accuracy was 0.64 or 64%, which we have achieved by the Support Vector Machine algorithm. Both of Which we can see from the diagrams (fig2 & fig3). Where we can see the precision, recall and f1-score. As we can see the best result with RandomForest classifiers was received by finding non-abusive data. The same was also seen with Support Vector Machine where best results were received from non-abusive data. One of the main reasons for this was that most of the data was neutral or non-abusive.

IV. RESULT ANALYSIS

The data set we have used for this experiment is 3006 sets of data. We have used the whole data set of 3006 in this experiment to determine the result of precision, recall,F-measure and accuracy. As we have used Term frequency-inverse document frequency and Count Vectorizerwe received two sets of results as well. With Term frequency-inverse document frequency the best accuracy we have achieved is with the Random Forest Algorithm, where the accuracy was 0.67 or 67% which we can see from table:3. As for the Count Vectorizer, from table 4, the best result of accuracy was 0.64 or 64%, which we have achieved by the Support Vector Machine algorithm. Both of Which we can see from the diagrams (fig2 & fig3). Where we can see the precision, recall and f1-score. As we can see the best result with Random Forest classifiers was received by finding non-abusive data. The same was also seen with Support Vector Machine where best results were received from non-abusive

data. One of the main reasons for this was that most of the data was neutral or non-abusive.

	precision	recall	f1-score	support
0	0.67	0.81	0.73	521
1	0.64	0.46	0.53	385
accuracy			0.66	906
macro avg	0.65	0.63	0.63	906
weighted avg	0.65	0.66	0.65	906

Fig. 2. RandomForestClassifier Using Tf_IDF

	precision	recall	f1-score	support
0	0.64	0.81	0.71	339
1	0.62	0.40	0.49	263
accuracy			0.63	602
macro avg	0.63	0.61	0.60	602
weighted avg	0.63	0.63	0.62	602

Fig. 3. Support Vector Machine(svm) using countvectorizer

CONCLUSION

In our research paper, we have experimented on abusive and Hate-speech detection of Bengali language. As we know there are very few works on this field in Bengali language so to achieve the best possible result, we have manually salvaged Bengali language data from social media and other platforms and we have manually done the annotation for better result and precision. We classified our data in two categories manually then ran our proposed algorithms on the data with two methods. We have also gone through a data cleaning process to reduce the garbage values and data from our actual dataset to purify our data. With our proposed algorithm we have received 64% accuracy at the beginning with the Count Vectorizer on SVM. But with the TF-IDF (Term frequency-inverse document frequency) vectorizer we achieved 67% with the Random Forest Algorithm. Although the result was great, there are still so many future possibilities from this result as the resources we currently had are limited. For starters with stemming or lemmatization, we could have increased the accuracy further. But due to unavailable resources in Bengali language we had to work without those. Therefore, a comparative analysis was not necessary in this case. As for the recent methods and techniques of machine learning, data requirements are very high for the best results and implementation, while the datasets we used in the research were not big enough to implement techniques and receive any greater results.

REFERENCES

[1] Emon, E.A., Rahman, S., Banarjee, J., Das, A.K. and Mittra, T., 2019, June. A deep learning approach to detect abusive Bengali text. In 20197th International Conference on Smart Computing Communications (ICSCC) (pp. 1-5). IEEE

[2] Warner, W. and Hirschberg, J., 2012, June. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19-26).

[3] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M. and Tesconi, M., 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (pp. 86-95

[4] Malmasi, S. and Zampieri, M., 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

[5] Das, A.K., Al Asif, A., Paul, A. and Hossain, M.N., 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, *30*(1), pp.578-591.

[6] Hussain, M.G., Al Mahmud, T. and Akthar, W., 2018, December. An approach to detect abusive bangla text. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)* (pp. 1-5). IEEE.

[7] Ishmam, A.M. and Sharmin, S., 2019, December. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 555-560). IEEE

[8] Jahan, M.S., 2020. CYBER BULLYING IDENTIFICATION ANDTACKLING USING NATURAL LANGUAGE PROCESSING TECH-NIQUES

[9] https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

[10] https://globalvoices.org/2020/04/05/bangladeshis-turn-to-video-sites-during-covid-19-lockdown/

[11] https://www.ethnologue.com/guides/ethnologue200

[12] https://www.mustgo.com/worldlanguages/bengali/

[13] Eshan, S.C. and Hasan, M.S., 2017, December. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.

[14] Liu, C.Z., Sheng, Y.X., Wei, Z.Q. and Yang, Y.Q., 2018, August. Research of text classification based on improved TF-IDF algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)* (pp. 218-222). IEEE.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

[16] https://github.com/MominulJM/Hate_Speech/tree/main