Shahzaib (18k-0272)

Muneeb (18k-1163)

HATe speech detection

Project Report

# **Abstract:**

As online content continues to grow, so does the spread of hate speech. Cyberbullying is increasing day by day and effecting many people life. Looking at all this we came up with the idea to classify text/tweets into 3 classes. We got the kaggle dataset for hatespeach containing tweets belonging to 3 classes (Hate, Offensive, Neither) all labeled with the count number of people classified them as (Hate, Offensive, Neither). We marked each tweet with the class which was selected by maximum people. We trained several models using that data and compared their scores. Tweets containing offensive words e.g. (F\*\*k you) etc. were classified in offensive class while simple day to day tweets were classified in neither class.

# **Introduction:**

The debate around the regulation of hate speech is still ongoing it is still not clear whether the best response to it is through legal measures, or other methods Regardless of the means of countering it, the evident harm of hate speech makes its detection crucial. Both the volume of content generated online, particularly in social media, and the psychological burden of manual moderation supports the need for the automatic detection of offensive and hateful content.

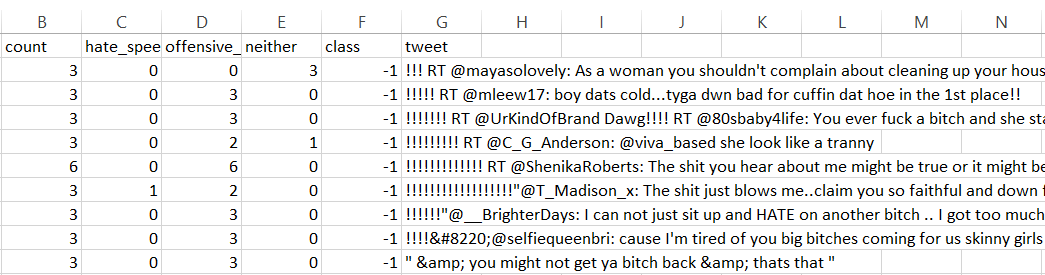
Hate speech is not at all easy to identify, hence the data we used was labeled as hate, offensive, or neither by individuals. That labels were used as benchmarks to train our model. The models we used were Random Forest, Naive Bayes, Logistic Regression, SVM and KNN

# **Problem Statement:**

In this project we aimed to build a model through recognizing patterns in the tweets that in turn can identify the offensive content on the social media and remove it so it doesn’t harm the society. Hate content being automatically detected and removed would make the social platform a lot cleaner and useable.

# **Dataset Description:**

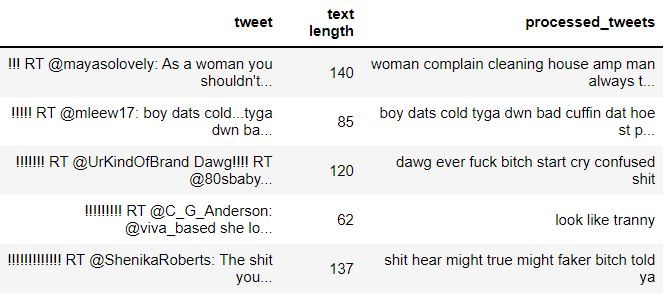
Dataset was taken from Kaggle which containing 27K random unprocessed tweets from all over the world , each tweet was marked as how many people classified it as either Hate, offensive or neither.



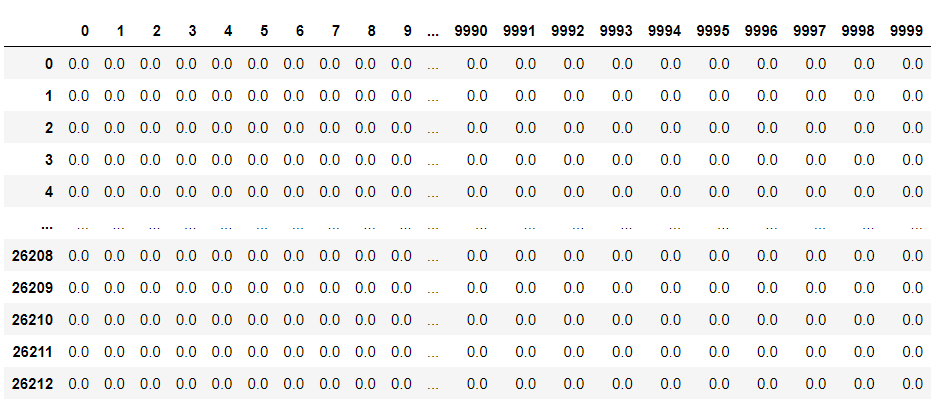
We marked the class label with respect to the max count of individuals marked that to a particular class. If 3 individuals marked a particular tweet as hate and only 1 marked a particular tweet as neither, that tweet was marked as hate.

# **Methodology:**

Starting with the dataset, tweets were in raw text form, containing retweets, links, special characters, stop words, white spaces, and @mentions. Tweets were first preprocessed i.e. all the stop words, special characters, white spaces, etc. were removed and were added to the dataset side by side with unprocessed tweets.

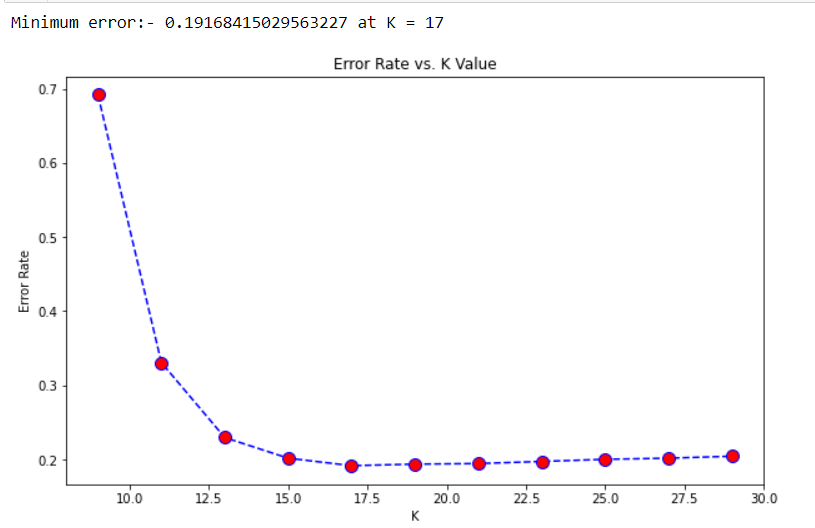


Then through TfidfVectorizer we converted those processed tweets into tokens and calculated it tf.idf score. Below a matrix is shown where each row is a vector or tweet and each column is a feature (token). Each cell represents the tf.idf score.



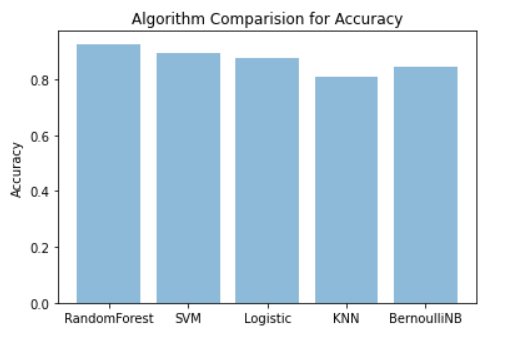
Then these vectors were split into the splits of 20% and 80% (training and testing data) several models were trained using these training vectors and then were tested using the testing data.

Several parameters were used in order to improve the accuracy and to make sure model doesn’t overfit or underfit. For knn we first found the K with the lowest error and then used that into our model.



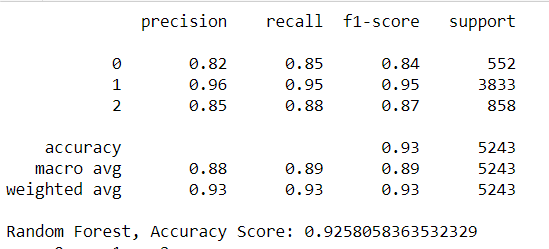
# **Result and Discussion:**

After applying out models to the dataset, we got different scores for each model.



RandomForestClassifier performing the best with overall accuracy of 0.92 and f1-score averaging 0.90 for all classes. With data of hate and offensive being highly overlapping this little error occurred due to some of the hate tweets classified as offensive and some offensive tweets as hate. Sometimes words like (f\*\*K) are used in day to language which usually is neither, with model not understanding the context of the sentence it classifies it either hate or offensive

Below shows the scores for Random forest Classifier.



# **Conclusion:**

This project was proved to be very challenging, sometimes tweets with actual label (hate) were classified as being offensive, the reasons were that data was bit biased towards offensive and other reason was that models can’t understand the context of the sentence. Sometimes tweets which were neither were classified as either hate or offensive just because it contained some slang words which are usually used in our day to day life.

# **Future Work:**

As hate speech continues to be a problem, the need for automatic hate speech detection systems becomes more apparent. We presented the current approaches for this task with our models achieving reasonable accuracy. But there is a huge room of improvement, in future we can make our model understand the context of the sentence. This proposed new approach can outperform existing systems at this task. Given all the challenges that remain, there is a need for more research on this problem. Our struggle to make internet safe for everyone continues.

# **References:**

1. <https://thecleverprogrammer.com/2020/08/19/hate-speech-detection-model/>
2. <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>
3. <https://matplotlib.org/>