VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

SOCIAL AND INFORMATION NETWORKS (CSE3021)
J COMPONENT FINAL REPORT

SENTIMENT ANALYSIS OF TWITTER TO DECIPHER AND DETECT OFFENSIVE SPEECH

Team Members:

Saurav Ranjan (19BCE0455)
Saksham Minocha (19BCE0466)

FACULTY: PROF. JAYA SUBALAKSHMI R

Domain:

Sentiment Analysis has become essential business wise as well as socially so as to analyze how millions of people take in the information and changes happening around the world and how it affects their lives. With growing popularity of social media and the anonymity and convenience it offers, has led to an increase in hate speech, therefore, there is an urgent need for effective solutions.

In our project we will perform sentiment analysis on twitter and detect the hate speech on tweets by using machine learning models. Twitter is an information network and communication mechanism that produces more than millions of tweets a day. The Twitter platform offers access to that corpus of data, via the APIs. Each API represents a facet of Twitter, and allows developers to build upon and extend their applications in new and creative ways. The Twitter API allows one to access the features of Twitter without having to go through the website interface. This can be useful for doing things like posting tweets or sending directed messages in an automated way with scripts. So, we will be using the twitter dataset which will be extracted from twitter API using an extractor and then carry out the methodologies and perform the ML algorithms.

Abstract:

Toxic online content has become a major issue in today's world due to an exponential increase in the use of the internet by people of different cultures and educational backgrounds. As and when cultures and backgrounds collide at a common standing, there is bound to be difference in opinions leading to debates and verbal clashes turning toxic at certain sensitive topics. There are around 500 million messages and posts created on the platform daily and thus the job of filtering toxic content and offensive language is a difficult task. Differentiating hate speech and offensive language is a key challenge in the automatic detection of toxic text content.

In this project, we plan to create a system using a learning approach to automatically classify tweets on Twitter into offensive speech and non-offensive. We will be using the twitter dataset which will be extracted from twitter API using an extractor and then we will perform the data-cleaning procedure - bag of words algorithm and the frequency-inverse document frequency (TFIDF) values to the machine learning model. We will then perform comparative analysis and study the effectiveness of the models considering both approaches.

Introduction:

The 21st century has seen various advancements in computing technology, but none comes close to being as effective as the discovery of machine learning. With its several advantages like pattern detection, no human intervention, continuous improvement, multi-dimensional and variety data support and flexible application, machine learning has come out to be the most sought-after technology there is today. The internet has experienced a mass influx of users that hail from different cultures and backgrounds. As and when cultures and backgrounds collide at a common understanding, there is bound to be difference in opinions leading to debates and verbal clashes turning toxic at certain sensitive topics. There are around 500 million messages and posts created on the platform daily and thus the job of filtering toxic content and offensive language is a difficult task. So, in this project, we plan to create a system using a learning approach to automatically classify tweets on Twitter into offensive speech and non-offensive.

Objective:

There are around 500 million messages and posts created on the platform daily and thus the job of filtering toxic content and offensive language is a difficult task. Due to the ongoing battles with hate speech which include dialogues related to sexism, racism, etc. we intend to decipher what can be classified to be hate speech and what cannot be.

- ➢ Our project aims to detect offensive speech in order to prevent the spreading of hate/toxic content over social media.
- ➢ Our project, thus will be helpful to act as a surveillance system so as to keep a check on the offensive messages/tweets on twitter.
- ➢ So, in this project, we plan to create a system using a learning approach to automatically classify tweets on Twitter into offensive speech and non-offensive.

Literature Survey:

Authors and Year (Reference)	Title (Study)	Features/Theoretical Model / Framework	Methodology Used/ Implementation	Dataset Details/ Analysis
Viswapriya, S. E., Gour, A. J. A. Y., & Gopi Chand, B. (2021, April)	Detecting Hate Speech and Offensive Language on Twitter using Machine Learning	N-gram, tf-idf, Regression, Naive-bayes, Support Vector Machine	Each model is trained on a training dataset by performing grid search for all the combinations of feature parameters and perform 10-fold cross-validation. The performance of these three algorithms is compared.	Crowd Flower (contains hateful, offensive, clean), Github (Tweet id, class)

Benching,W., Gupta, A., Vavor, A., Li, J., Umair, H., Durzynski,N. (2021, August)	Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning	Word Embeddings, Bi-directional LSTM, BERT-Model, Hugging Face auto tokenizer	A transfer learning approach for hate speech detection using an existing pre-trained language model BERT (Bidirectional Encoder Representations from Transformers), DistilBert (Distilled version of BERT) and GPT-2 (Generative Pre-Training) is introduced along with hyper parameters tuning analysis.	Public Tweet Data comprising 24,783 tweets with 6 columns: count, hate-speech, offensive-language, neither, class, and tweet.

Lizhou Fan, Huizi Yu, and Zhanyuan Yin (2020, October)	Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter	Decision tree classifier, Network methods	Utilizing data visualization as an exploratory research method and effectively identifying patterns and correlation among variables and Hashtag trends. After removing stop words and ignoring non-textural strings, they used lexicon-based methods for the primary data wrangling of both Hate Speech detection and aspect-based emotion scoring.	Database of tweets, External data collection
Ricardo Martins, Marco Gomes, Jos´e Jo˜ao Almeida, Paulo Novais, Pedro Henriques (2018, December)	Hate speech classification in social media using emotional analysis	NLP,TF-IDF	The discrete emotional models consider that every emotion is composed of universally displayed and recognized basic emotions, as happiness, anger, sadness, surprise, disgust, and fear, for instance.	Dataset from Kaggle by Davidson and Warmley containing 24782 tweets.

Yiwen Tang and Nicole Dalzell (2019, August)	Classifying Hate Speech Using a Two-Layer Model	Two-Layer Model	One of the main advantages of discrete models is that, through psychophysical experiments, the perception of emotions by human beings is discrete.	Wikipedia
Zia, T., Akram, M. S., Nawaz, M. S., Shahzad, B., Abdullatif, A. M., Mustafa, R. U., & Lali, M. I. (2017, June)	Identification of hatred speeches on Twitter	unigrams, TF-IDF, retweets, favourites, page autenticity	Used supervised learning algorithms and selected three widely used algorithms SVM, NB and kNN. These algorithms were used to classify religious opinions first and then found their sentiment. Founded SVM as best classifier for sentiment classification	tweets

Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. (2018, June)	Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media	N grams, semantic and syntactic, T F-IDF, word2vec embeddings, doc2vec embeddings	Main contribution is two-fold: first, the granular taxonomy of hateful online comments. Identified categories and subcategories of hateful speech from the social media comments, forming a comprehensive taxonomy for machine learning. Second, trained a multiclass, multilabel model that classifies the hateful comments, experimenting with Logistic Regression, Decision Tree, Random Forest, Adaboost, and SVM.	5143 labeled comments YouTube and Facebook videos
Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April)	Deep learning for hate speech detection in tweets	char ngrams, TFIDF, BoWV, rand om embeddings, GloVe embeddings	Experimented with three broad representations. (1) Char n-grams: It is the state of-the-art method which uses character n-grams for hate speech detection (2) TF-IDF: TF-IDF are typical features used for text classification. (3) BoWV: Bag of Words Vector approach uses the average of the word (GloVe)	16,914 annotated tweets

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May)	Automated hate speech detection and the problem of offensive language	n-grams, TFIDF, POS, readability, sentiment, hashtags, mentions, retweets, URLs, length	Used a logistic regression with L1 regularization to reduce the dimensionality of the data. Tested each model using 5-fold cross validation, holding out 10% of the sample for evaluation to help prevent overfitting. Decided to use a logistic regression with L2 regularization for the final model. Used a one-versus-rest framework where a separate classifier is trained for each class and the class label with the highest predicted probability across all classifiers is assigned to each tweet. All modelling was performed using scikit-learn	24,802 labeled tweets.

Gambäck, B., & Sikdar, U. K. (2017, August)	Using convolutional neural networks to classify hate speech	CNN	Two CNN models were created based on different input vectors sets that were fed to the neural networks for training and classification. Word vectors based on semantic information were built using an unsupervised strategy, word2vec, and compared to a randomly generated vector baseline. In additional, two CNN models were trained on character 4-grams, as well as on a combination of word vec	6655 tweets from annotated tweets

List of Modules:

1.Pre-Processing

a) Bag of Words

A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The objective is to turn each document of free text into a vector that we can use as input or output for a machine learning model. Because we know the vocabulary has 10 words, we can use a fixed-length document representation of 10, with one position in the vector to score each word. The simplest scoring method is to mark the presence of words as a Boolean value, 0 for absent, 1 for present.

b) TFIDF

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term. Put simply, the higher the TF*IDF score (weight), the rarer the term and vice versa.
It is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the web, which is referred to as corpus. For a term t in a document d, the weight Wt, d of term t in document d is given by: Wt, d = TFt, d log (N/DFt)
Where:

- TFt, d is the number of occurrences of t in document d.
- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

(2) Data Cleaning

a) Stop Words Removal
Stop words are usually articles or prepositions which do not help us to find the context or the true meaning of a sentence. These are words that can be removed without any negative consequences to the final model that you are training. Commonly used in English language include "is"," and"," are" etc.

b) Greek Words Removal
We next removed special characters and numbers. Numbers rarely contain useful meaning. Special characters can bloat our term-frequency matrix.

c) Slang Words
Slang words include informal short forms of words which are usually used in speech. Due to lack of an existing dictionary, we mapped a few slang words to their original forms such as "luv" to love, "thx" to thanks to mention a few examples.

d) Stemming
With stemming, words are reduced to their word stems by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word. A word is looked at and run through a series of conditionals that determine how to cut it down. For our dataset, the porter stemming algorithm was used .

e) Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meaning to one word. For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.

3. Training Modules

a) Gradient Boosting

A greedy algorithm and can overfit a training dataset quickly. It can benefit from regularization methods that penalize various parts of the algorithm and generally improve the performance of the algorithm by reducing overfitting

b) Decision Tree

A Supervised learning technique that can be used for both classification and Regression problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches

c) Naive Bayes

A supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

d) Random Forest

It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output

e) Logistic Regression

One of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.Logistic regression is used for solving the classification problems.

Modules Flowchart:

The architecture with the respective modules are layered out in the flowchart below:

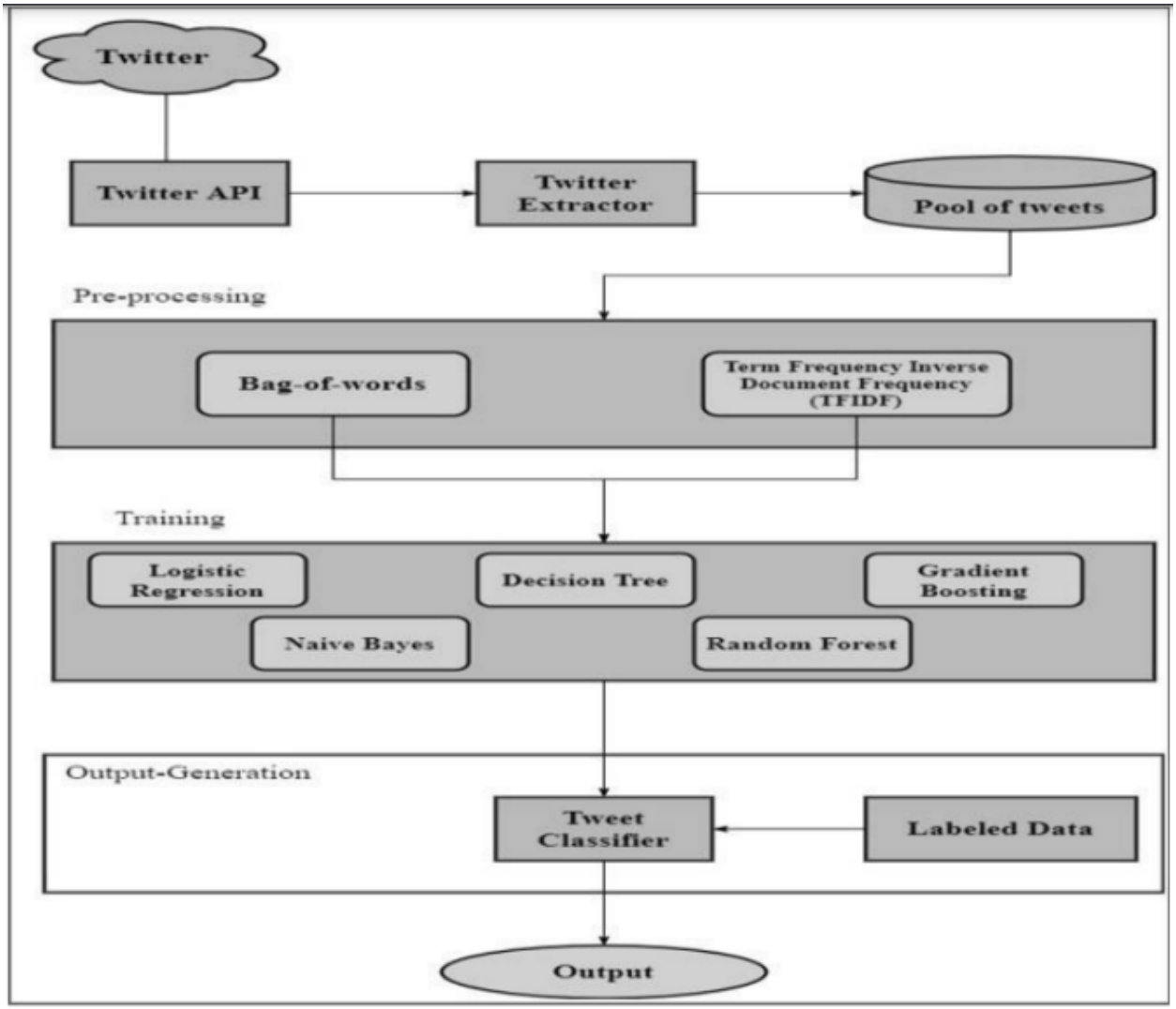


Fig: Modules Flowchart

Data Architecture:

The data we have is imbalanced. We have a total of 31935 rows out of which the hate tweets comprise only 7% of the total tweets. Modeling on an imbalanced dataset is not ideal since the model won't be able to learn what pieces of text or information causes the tweet to be classified as a hate tweet. If we train a model on imbalanced data, the results will be misleading. In such datasets, due to lack of learning, the model would simply predict each tweet as a good tweet.
In this case, in spite of having a high accuracy, it is not actually doing a good job of classification since it is unable to classify the hate tweets accurately. Therefore, we perform strategic sampling and separate the data into a temporary set and test set. Note that since we have performed strategic sampling, the ratio of good tweets to hate tweets is 93:7 for both the temporary and test datasets. On the temporary data, we first tried to perform a sampling of hate tweets using SMOTE (Synthetic Minority Oversampling Technique). Since the SMOTE packages don't work directly for textual data, we wrote our own code for it.
The process is as follows:
We created a corpus of all the unique words present in hate tweets of the temporary dataset. Once we had a matrix containing all possible words in hate tweets, we created a blank new dataset and started filling it with new hate tweets. These new tweets were synthesized by selecting words at random from the corpus. The lengths of these new tweets were determined on the basis of the lengths of the tweets from which the corpus was formed.
We then repeated this process multiple times until the number of hate tweets in this synthetic data was equal to the number of non-hate tweets we had in our temporary data. However, when we employed the Bag of Words approach for feature generation, the number of features went up to 100,000. Due to an extremely high number of features, we faced hardware and processing power limitations and hence had to discard the SMOTE oversampling method.
As it was not possible to up-sample hate tweets to balance the data, we decided to down-sample non-hate tweets to make it even. We took a subset of only the non-hate tweets from the temporary dataset. From this subset, we selected n random tweets, where n is the number of hate tweets in the temporary data. We then joined this with the subset of hate tweets in the temporary data. This dataset is now the training data that we use for our feature generation and modeling purposes. The test data is still in a 93:7 ratio of good tweets to hate tweets as we did not perform any sampling on it. Sampling was not performed as real world data comes in this ratio.

Methodology

1. The first step of building our model was to balance the number of hate and non-hate tweets.
2. We clean the tweets by employing lemmatization, stemming, removal of stop words, and omissions.
3. Then for the pre-processing step, we use Bag of words and Term Frequency Inverse Document Frequency (TFIDF).

4. For both the Bag of words and TFIDF, we run 5 classification algorithms, namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest and Gradient Boosting.
5. These 5 algorithms are run again after performing dimensionality reduction for both TF-IDF and Bag of Words.

Our goal is to classify tweets into two categories, hate speech or non-hate speech. Our project analyzed a dataset CSV file from Kaggle containing 31,935 tweets. The dataset was heavily skewed with 93% of tweets or 29,695 tweets containing non-hate labeled Twitter data and 7% or 2,240 tweets containing hate-labeled Twitter data. The first step of building our model was to balance the number of hate and non-hate tweets. Our data preprocessing step involved 2 approaches, Bag of words and Term Frequency Inverse Document Frequency (TFIDF). The bag-of-words approach is a simplified representation used in natural language processing and information retrieval. In this approach, a text such as a sentence or a document is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. TFIDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is used as a weighting factor in searches of information retrieval, text mining, and user modeling. Before we input this data into various algorithms, we have to clean it as the tweets contain many different tenses, grammatical errors, unknown symbols, hashtags, and Greek characters.

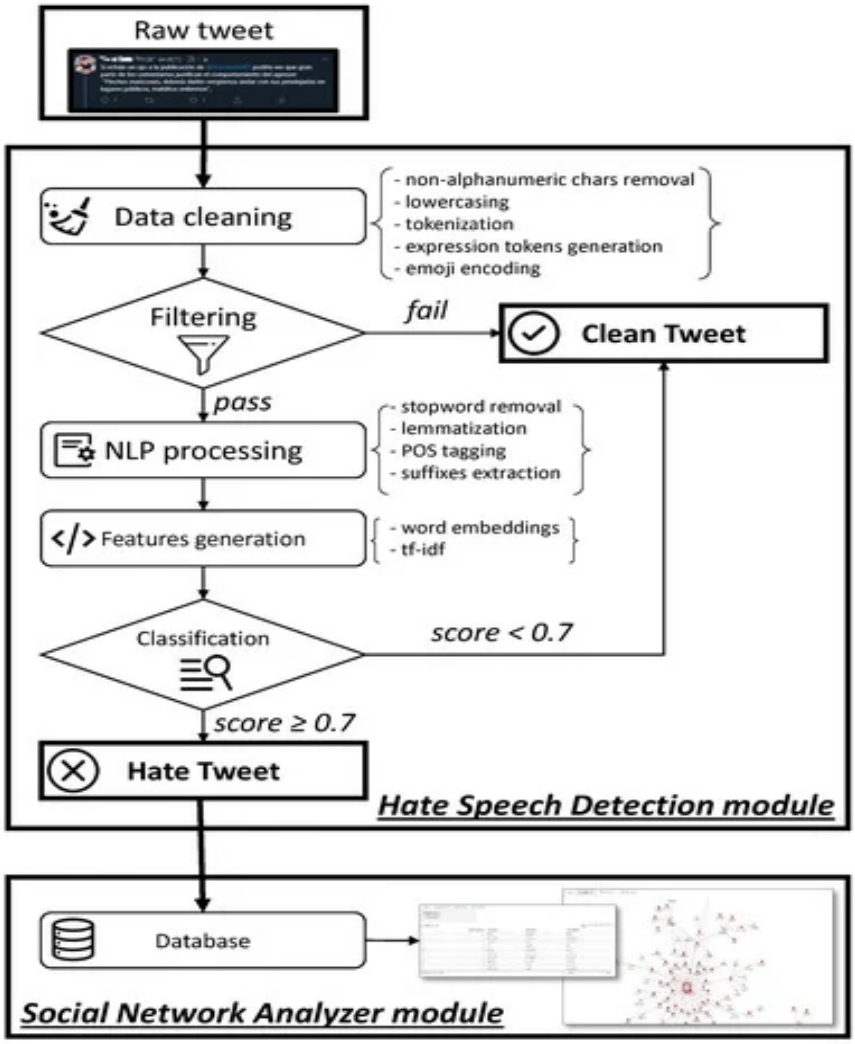
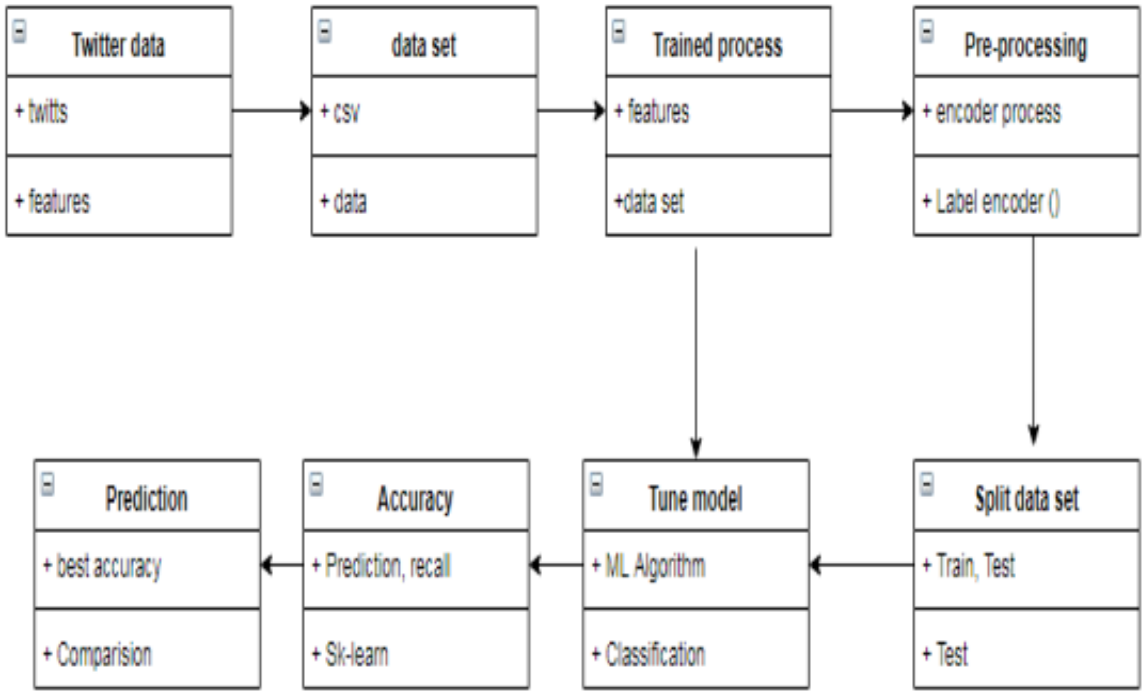


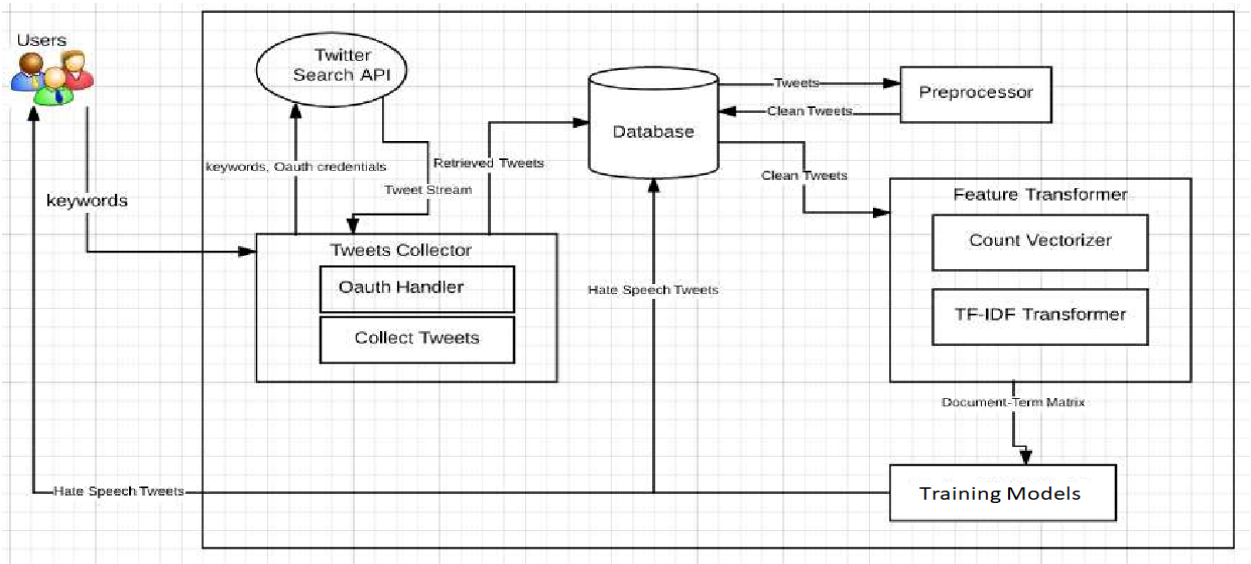
Fig: Architecture Diagram

UML Diagrams:

1. Class Diagram:



2. Use Case Diagram:



Hardware Requirements:

- CPU: Intel® Core™ CPU i3 or Higher
- GPU: Intel HD Graphics or Higher
- Memory: 4 GB, DD4 or Higher
- Storage: 250GB or Higher

Software Requirements:

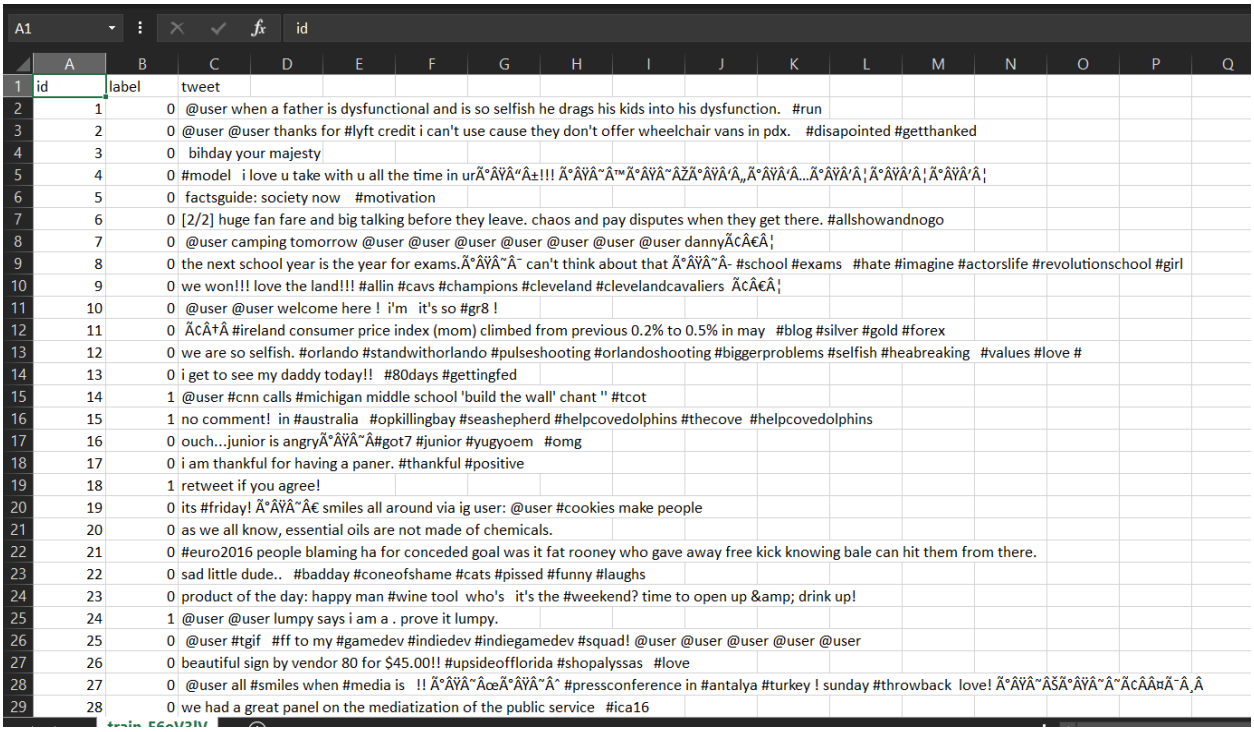
- Any Operating System
- Python 3
- Vs Code
- Jupyter
- Google Collab

Libraries Used:

- **NumPy :** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- **Pandas :** pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.
- **Scikit-learn :** Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- **NLTK :** The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.
- **Matplotlib :** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or G.

Dataset Used:

Our project analyzed a dataset CSV file from Kaggle containing 31,935 tweets. The dataset was heavily skewed with 93% of tweets or 29,695 tweets containing non-hate labeled Twitter data and 7% or 2,240 tweets containing hate-labeled Twitter data.



	id	label	tweet
2	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
3	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
4	3	0	bihday your majesty
5	4	0	#model i love u take with u all the time in urÃ°ÂŸÂ“Â±!!! Ã°ÂŸÂ™Ã°ÂŸÂ˜ÂŽÃ°ÂŸÂ‘Â„Ã°ÂŸÂ‘Â…Ã°ÂŸÂ‘Â¦Ã°ÂŸÂ‘Â¦Ã°ÂŸÂ‘Â¦
6	5	0	factsguide: society now #motivation
7	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
8	7	0	@user camping tomorrow @user @user @user @user @user @user dannyÃcÂ€Â¦
9	8	0	the next school year is the year for exams.Ã°ÂŸÂ˜Â¯ can't think about that Ã°ÂŸÂ˜Â- #school #exams #hate #imagine #actorslife #revolutionschool #girl
10	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers ÃcÂ€Â¦
11	10	0	@user @user welcome here ! i'm it's so #gr8 !
12	11	0	ÃcÂ†Â #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
13	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
14	13	0	i get to see my daddy today!! #80days #gettingfed
15	14	1	@user #cnn calls #michigan middle school 'build the wall' chant '' #tcot
16	15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
17	16	0	ouch...junior is angryÃ°ÂŸÂ˜Â#got7 #junior #yugyoem #omg
18	17	0	i am thankful for having a paner. #thankful #positive
19	18	1	retweet if you agree!
20	19	0	its #friday! Ã°ÂŸÂ˜Â€ smiles all around via ig user: @user #cookies make people
21	20	0	as we all know, essential oils are not made of chemicals.
22	21	0	#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick knowing bale can hit them from there.
23	22	0	sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs
24	23	0	product of the day: happy man #wine tool who's it's the #weekend? time to open up & drink up!
25	24	1	@user @user lumpy says i am a . prove it lumpy.
26	25	0	@user #tgif #ff to my #gamedev #indiedev #indiegamedev #squad! @user @user @user @user @user
27	26	0	beautiful sign by vendor 80 for $45.00!! #upsideofflorida #shopalyssas #love
28	27	0	@user all #smiles when #media is !! Ã°ÂŸÂ˜ÂœÃ°ÂŸÂ˜Âˆ #pressconference in #antalya #turkey ! sunday #throwback love! Ã°ÂŸÂ˜ÂŠÃ°ÂŸÂ˜ÃcÂÂ¤Ã¯Â¸Â
29	28	0	we had a great panel on the mediatization of the public service #ica16

	A	B	C	D	E	F	G
1		count	hate_speech	offensive_language	neither	class	tweet
2	0	3	0	0	3	2	!!! RT @mayasolove
3	1	3	0	3	0	1	!!!!! RT @mleew17: I
4	2	3	0	3	0	1	!!!!!!! RT @UrKindOfl
5	3	3	0	2	1	1	!!!!!!!!! RT @C_G_An
6	4	6	0	6	0	1	!!!!!!!!!!!! RT @Shenl
7	5	3	1	2	0	1	!!!!!!!!!!!!!!!!!"@T_Ma
8	6	3	0	3	0	1	!!!!!"@__BrighterDay
9	7	3	0	3	0	1	!!!!“@selfiequ
10	8	3	0	3	0	1	" & you might n
11	9	3	1	2	0	1	" @rhythmlxx_ :hobb bitch
12	10	3	0	3	0	1	" Keeks is a bitch she
13	11	3	0	3	0	1	" Murda Gang bitch
14	12	3	0	2	1	1	" So hoes that smoke
15	13	3	0	3	0	1	" bad bitches is the
16	14	3	1	2	0	1	" bitch get up off me
17	15	3	0	3	0	1	" bitch nigga miss m
18	16	3	0	3	0	1	" bitch plz whatever
19	17	3	1	2	0	1	" bitch who do you l
20	18	3	0	3	0	1	" bitches get cut off
21	19	3	0	3	0	1	" black bottle &
22	20	3	0	3	0	1	" broke bitch cant te
23	21	3	0	3	0	1	" cancel that bitch li
24	22	3	0	3	0	1	" cant you see these
25	23	3	0	3	0	1	" fuck no that bitch c
26	24	3	0	3	0	1	" got ya bitch tip toe
27	25	3	0	2	1	1	" her pussy lips like H

Results and Discussions:

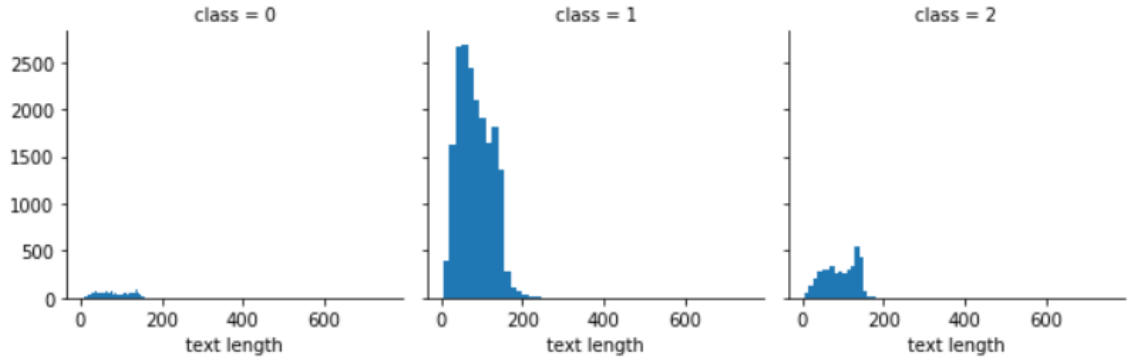
```
1 df.head()
```

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

Displaying the information as obtained from the dataset

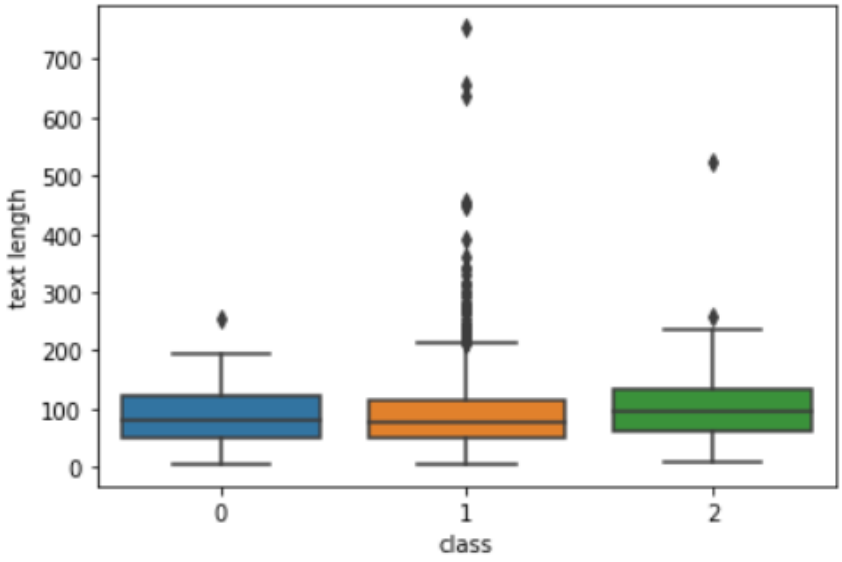
```
[7]  1 #Basic visualization of data using histograms
     2 # FacetGrid- Multi-plot grid for plotting conditional relationships
     3 import seaborn as sns
     4 import matplotlib.pyplot as plt
     5 graph = sns.FacetGrid(data=dataset, col='class')
     6 graph.map(plt.hist, 'text length', bins=50)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f7e2a027550>
```




```
[8]  1 # Box-plot visvualization
     2 sns.boxplot(x='class', y='text length', data=dataset)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7e27c81810>
```



	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause they...
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ur...
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before the...	[2/2] huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...	camping tomorrow dannya|
7	8	0	the next school year is the year for exams.ð□□...	the next school year is the year for exams.d- ...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	we won!!! love the land!!! #allin #cavs #champ...
9	10	0	@user @user welcome here ! i'm it's so #gr...	welcome here ! i'm it's so #gr8 !

Displaying the data after removing the missing values and removing "@"

	id	label	tweet	clean_tweet	Hash words
0	1	0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...	#run
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause they...	#lyft #disapointed #getthanked
2	3	0	bihday your majesty	bihday your majesty	No hashtags
3	4	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ur...	#model
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation	#motivation

Column showing whether the corresponding tweet has a hash tagged word or not

	id	label	tweet	clean_tweet	#	clean_tweet_final
0	1	0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...	#run	when a father is dysfunctional and is so selfi...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause they...	#lyft #disapointed #getthanked	thanks for credit i can't use cause they don't...
2	3	0	bihday your majesty	bihday your majesty		bihday your majesty
3	4	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ur...	#model	i love u take with u all the time in urd+-!!! ...
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation	#motivation	factsguide: society now
5	6	0	[2/2] huge fan fare and big talking before the...	[2/2] huge fan fare and big talking before the...	#allshowandnogo	[2/2] huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...	camping tomorrow dannya|		camping tomorrow dannya|
7	8	0	the next school year is the year for exams.ð□□...	the next school year is the year for exams.d- ...	#school #exams #hate #imagine #actorslife #rev...	the next school year is the year for exams.d- ...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	we won!!! love the land!!! #allin #cavs #champ...	#allin #cavs #champions #cleveland #clevelandc...	we won!!! love the land!!! a|
9	10	0	@user @user welcome here ! i'm it's so #gr...	welcome here ! i'm it's so #gr8 !	#gr8	welcome here ! i'm it's so !
10	11	0	â□□ #ireland consumer price index (mom) climb...	a #ireland consumer price index (mom) climbed ...	#ireland #blog #silver #gold #forex	a consumer price index (mom) climbed from prev...
11	12	0	we are so selfish. #orlando #standwithorlando ...	we are so selfish. #orlando #standwithorlando ...	#orlando #standwithorlando #pulseshooting #orl...	we are so selfish.
12	13	0	i get to see my daddy today!! #80days #getti...	i get to see my daddy today!! #80days #gettingfed	#80days #gettingfed	i get to see my daddy today!!
13	14	1	@user #cnn calls #michigan middle school 'buil...	#cnn calls #michigan middle school 'build the ...	#cnn #michigan #tcot	calls middle school 'build the wall' chant "

Obtaining the tweets after performing the data clean operation on the tweets

```
[ ]  1 #Tokenization
     2 corpus = []
     3 for i in range(0,31962):
     4     tweet = data_frame['clean_tweet'][i]
     5     tweet = tweet.lower()
     6     tweet = tweet.split()
     7     tweet = [ps.stem(word) for word in tweet if not word in set(stopwords.words('english'))]
     8     tweet = ' '.join(tweet)
     9     corpus.append(tweet)
```

```
[ ]  1 #Ensuring all the tweets are tokenized into individual words
     2 len(corpus)
```

```
31962
```

Performing the tokenization of the tweets

```
1 corpus
```

```
['father dysfunct selfish drag kid dysfunction. #run',
 "thank #lyft credit can't use cau offer wheelchair van pdx. #disapoint #getthank",
 'bihday majesti',
 '#model love u take u time urd+-!!! dddd d|d|d|',
 'factsguide: societi #motiv',
 '[2/2] huge fan fare big talk leave. chao pay disput get there. #allshowandnogo',
 'camp tomorrow dannya|',
 "next school year year exams.d- can't think #school #exam #hate #imagin #actorslif #revolutionschool #girl",
 'won!!! love land!!! #allin #cav #champion #cleveland #clevelandcavali a|',
 "welcom ! i'm #gr8 !",
 '#ireland consum price index (mom) climb previou 0.2% 0.5% may #blog #silver #gold #forex',
 'selfish. #orlando #standwithorlando #pulseshoot #orlandoshoot #biggerproblem #selfish #heabreak #valu #love #',
 'get see daddi today!! #80day #gettingf',
 "#cnn call #michigan middl school 'build wall' chant '' #tcot",
 'comment! #australia #opkillingbay #seashepherd #helpcovedolphin #thecov #helpcovedolphin',
 'ouch...junior angryd#got7 #junior #yugyoem #omg',
 'thank paner. #thank #posit',
 'retweet agree!',
 '#friday! smile around via ig user: #cooki make peopl',
 'know, essenti oil made chemicals.',
 '#euro2016 peopl blame ha conc goal fat rooney gave away free kick know bale hit there.',
 'sad littl dude.. #badday #coneofsham #cat #piss #funni #laugh',
 "product day: happi man #wine tool who' #weekend? time open &amp;amp; drink up!",
 'lumpi say . prove lumpy.',
 '#tgif #ff #gamedev #indiedev #indiegamedev #squad!',
 'beauti sign vendor 80 $45.00!! #upsideofflorida #shopalyssa #love',
 '#smile #media !! dd #pressconf #antalya #turkey ! sunday #throwback love! dda$?i,',
 'great panel mediat public servic #ica16',
 "happi father' day dddd",
```

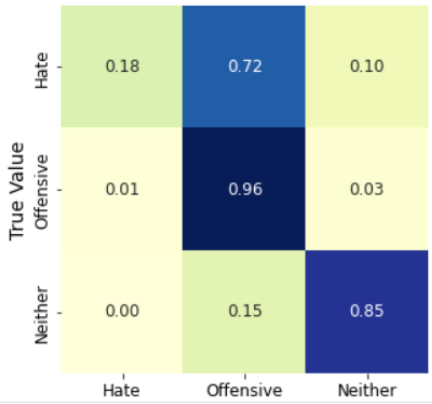
Tweets after being tokenized

```
[22]  1 # F2-Conctaenation of tf-idf scores and sentiment scores
      2 tfidf_a = tfidf.toarray()
      3 modelling_features = np.concatenate([tfidf_a,final_features],axis=1)
      4 modelling_features.shape
```

```
(24783, 6754)
```

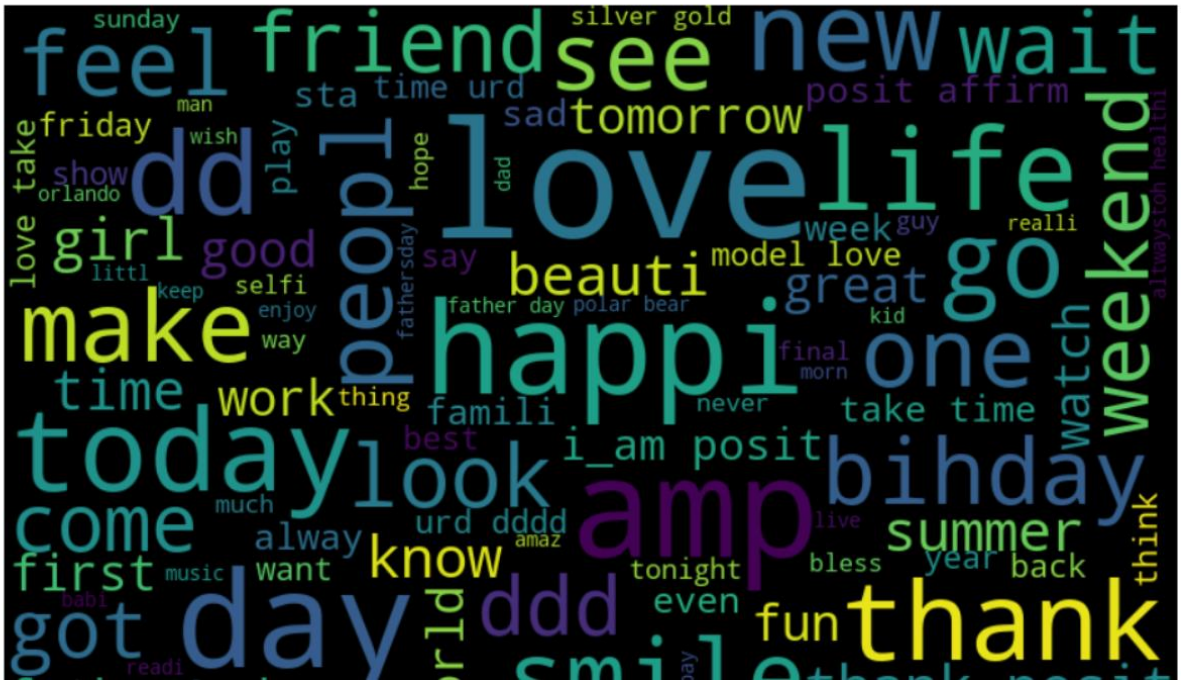
Concatenation of tf-idf scores

```
[ ]  1 #Confusion Matrix for TFIDF with additional features
     2 from sklearn.metrics import confusion_matrix
     3 confusion_matrix = confusion_matrix(y_test,y_preds)
     4 matrix_proportions = np.zeros((3,3))
     5 for i in range(0,3):
     6     matrix_proportions[i,:] = confusion_matrix[i,:]/float(confusion_matrix[i,:].sum())
     7 names=['Hate','Offensive','Neither']
     8 confusion_df = panda.DataFrame(matrix_proportions, index=names,columns=names)
     9 plt.figure(figsize=(5,5))
    10 seaborn.heatmap(confusion_df,annot=True,annot_kws={"size": 12},cmap='YlGnBu',cbar=False, square=True,fmt='.2f')
    11 plt.ylabel(r'True Value',fontsize=14)
    12 plt.xlabel(r'Predicted Value',fontsize=14)
    13 plt.tick_params(labelsize=12)
```



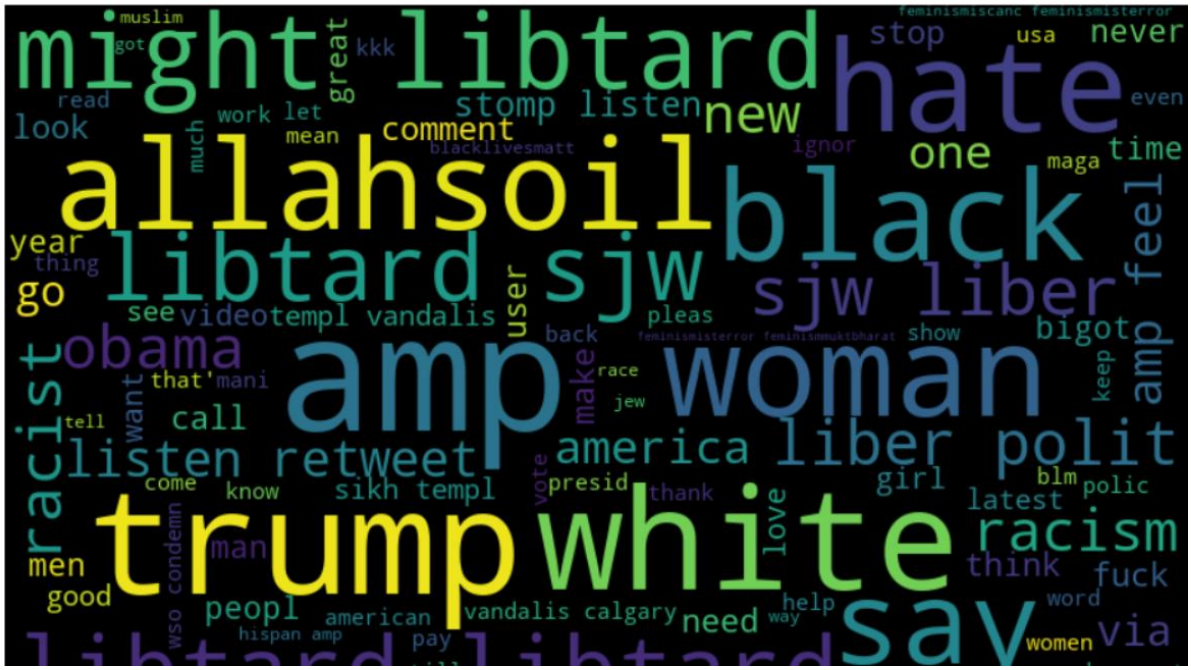
Confusion matrix for Tf-idf with additional features

Normal words
(-0.5, 799.5, 499.5, -0.5)

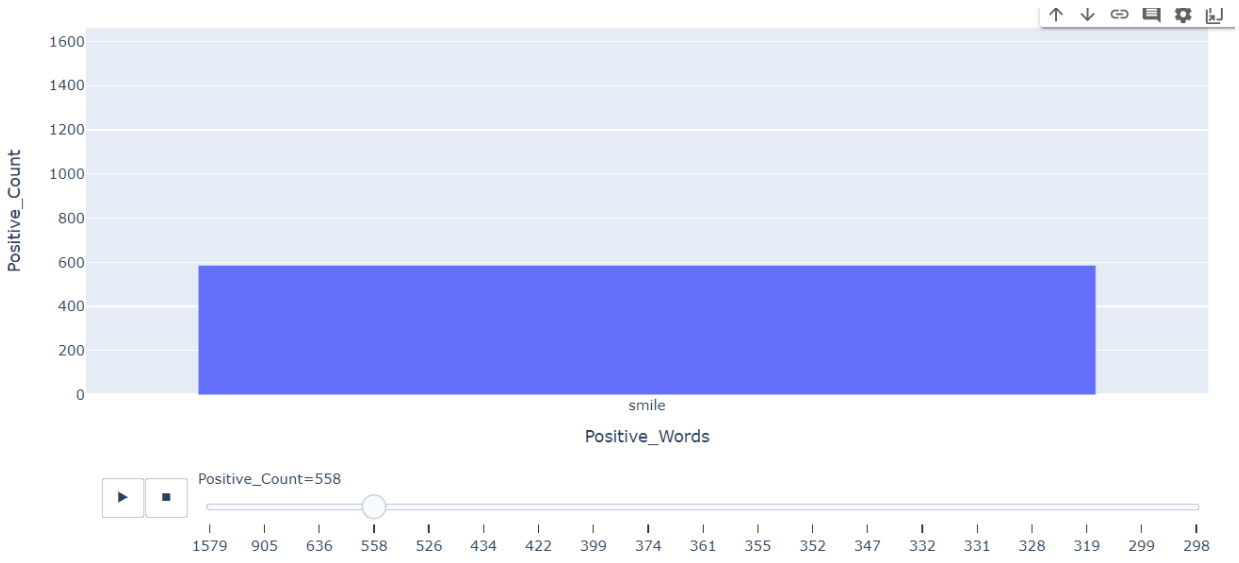


Word cloud for the most common words in positive tweets

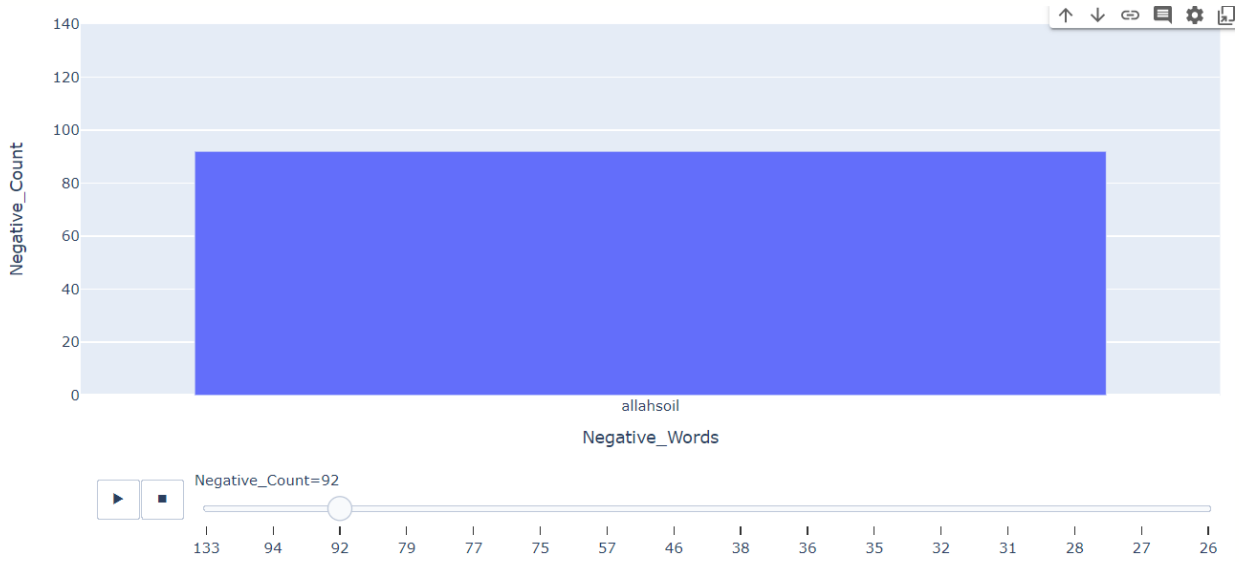
Normal words
(-0.5, 799.5, 499.5, -0.5)



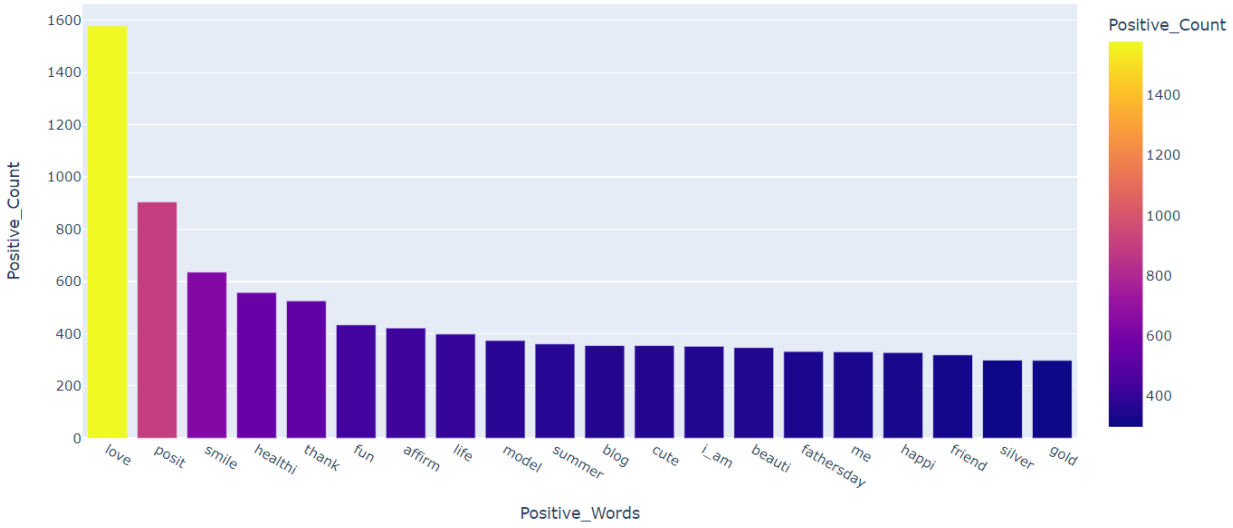
Word cloud for the most common words in negative tweets



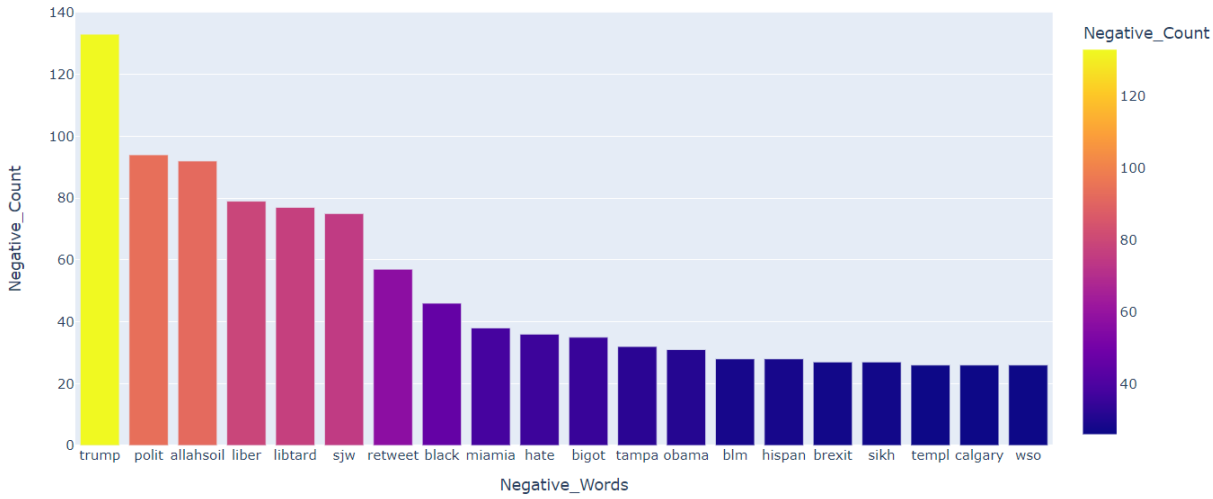
Animated plot for positive words for their frequency



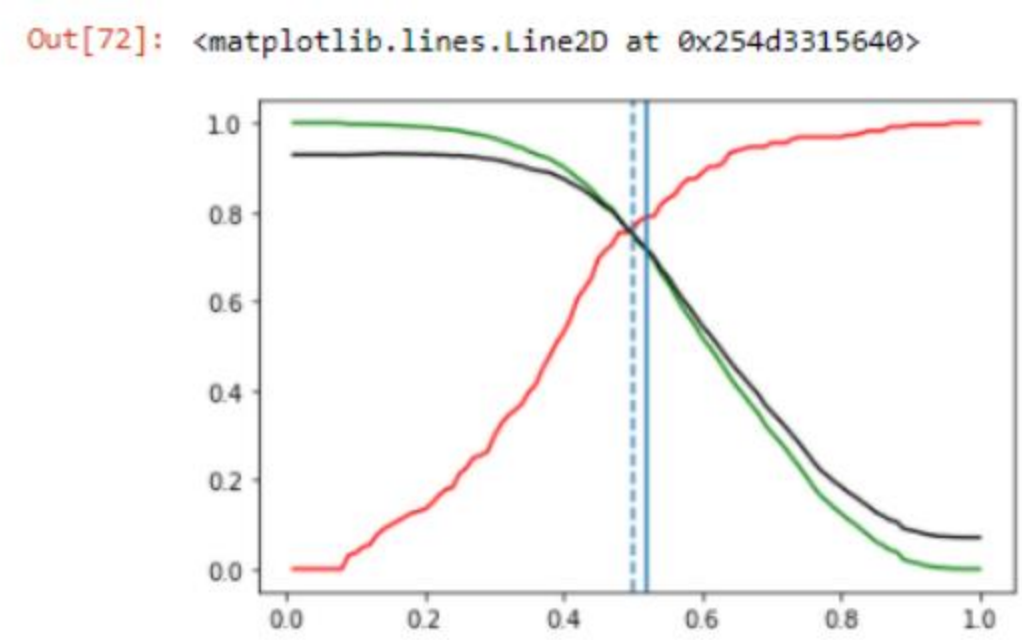
Animated plot for negative words for their frequency



Normal histogram of positive words



Normal histogram for negative words



TF-IDF Approach: Graph showing True Positive Rate(red), True Negative Rate(green) and overall accuracy(black) for TFIDF

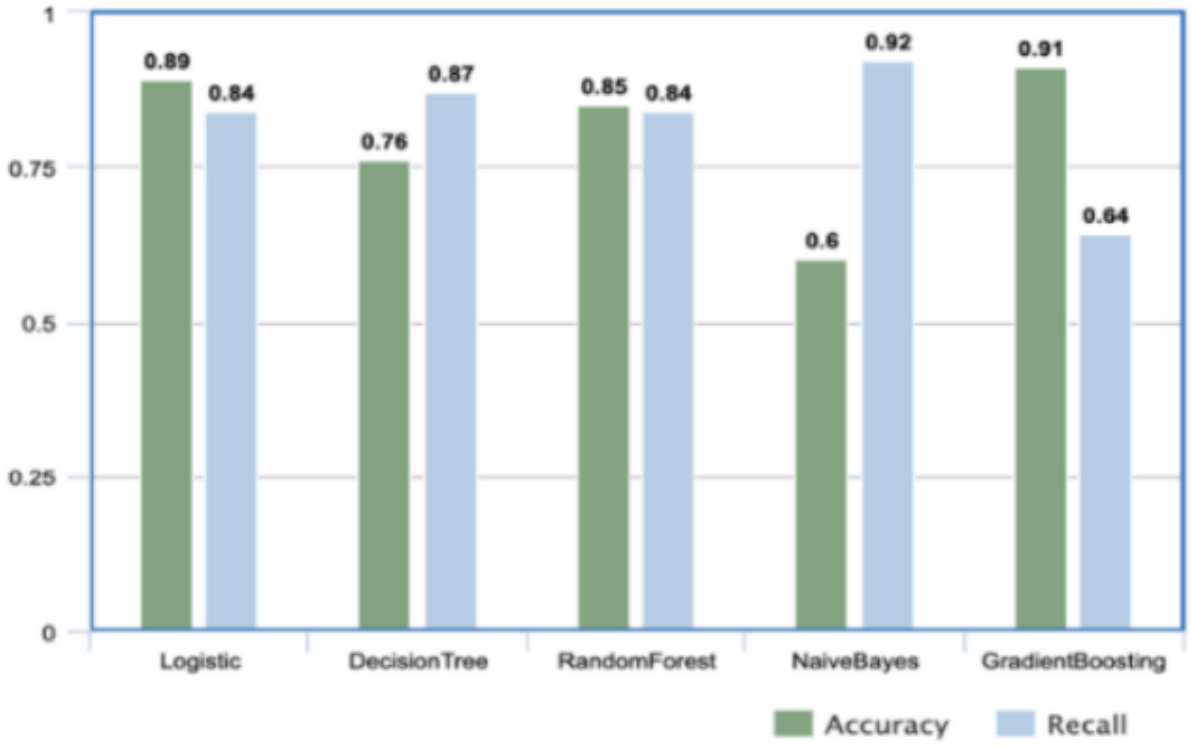
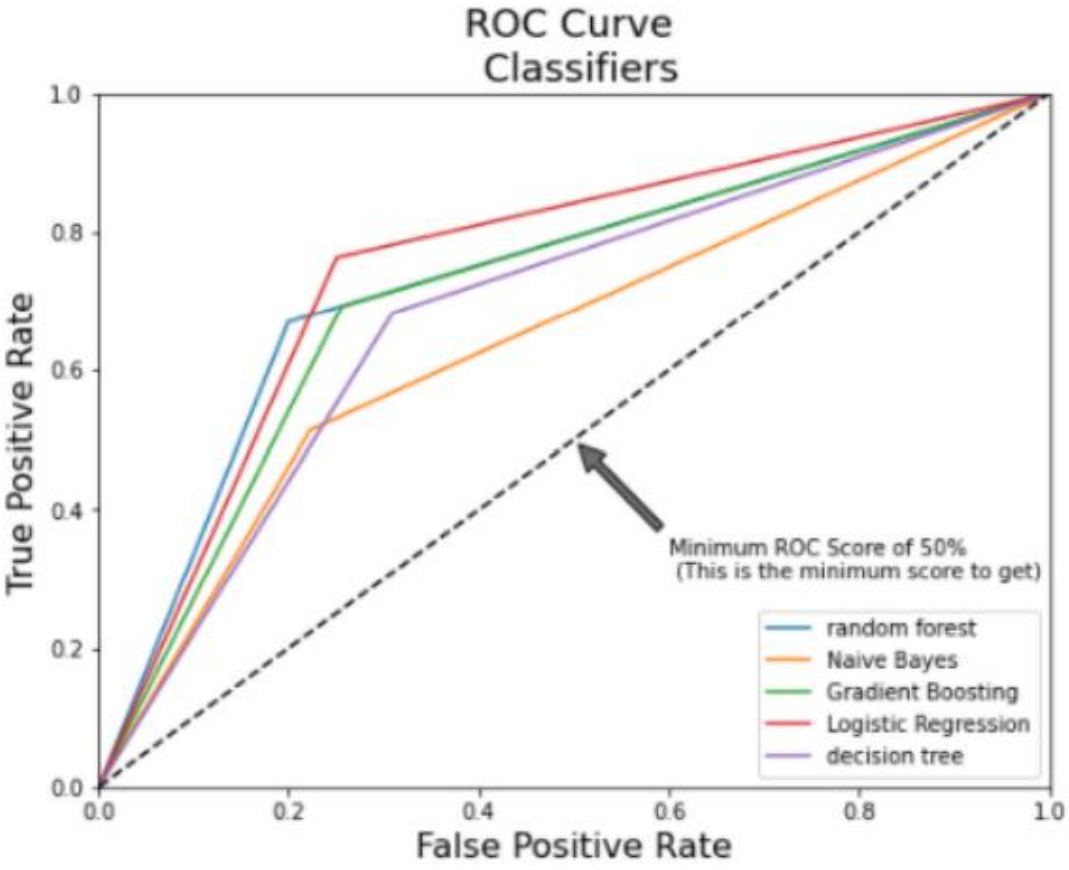


Chart representing Accuracy and recall of the cleaned tweets using TFIDF Approach

At a first glance, gradient boosting seems to be the best model due to its highest accuracy. But as discussed in data architecture, our test data contains a 93:7 ratio for non-hate tweets to hate tweets. Thus, having high accuracy may not be the best metric. We hence take into consideration recall, which is the proportion of hate tweets we are able to correctly identify. If this metric is high, it would correspond to a high number of hate tweets being classified as hate tweets and thus achieving our objective. If we look exclusively at recall, Naive Bayes seems to be the best model. However, another factor to be considered is that all the tweets we capture as hate tweets will be forwarded to an operations team which would review each of these tweets individually. Therefore, having a large number of false positives will mean that the analysts will have to unnecessarily scrutinize a higher number of tweets which are actually non hate tweets. This will lead to inefficiencies in the utilization of time, resources and personnel. It would be ideal to move forward with models which are giving high scores for both accuracy and recall. Thus, we select logistic regression and random forest as the champion models.



ROC Curve of the different classifiers in the TFIDF Approach

Algorithm	AUC
Logistic Regression	0.868
Naive Bayes	0.749
Random forest	0.848
Decision Tree	0.811
Gradient Boosting	0.789

Area Under Curve of the ROC Curve for the classifiers

Here is the AUC score for the classifiers. We can see out of the 5 Logistic Regression performs the best. For ROC the curve score should be near to 1 and it is close to 1 (0.869). Naïve Bayes classifier which uses probabilistic determination is the worst out of all with a low score.

Conclusion

As we can see, logistic regression gives us the best AUC. Also, if we look at its ROC curve, we can see that it covers the False Positive Rate (FPR) and True Positive Rate (TPR) area equally. Compare this to Naive Bayes which shows a poor false positive performance, which was also evident by its accuracy score earlier. Similarly, gradient boosting has a poor true positive performance which was evident by its low recall score. In addition, logistic regression is straightforward in nature, i.e. the model can be explained with the help of a simple exponential equation, and its coefficients can be visualized and compared. Therefore, logistic regression comes out as the best model for our case.

References:

[1]Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. (2021). Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning. arXiv preprint arXiv:2108.03305.

[2] Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651.

[3] Fan, L., Yu, H., & Yin, Z. (2020). Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. Proceedings of the Association for Information Science and Technology, 57(1), e313.

[4] Tang, Y., & Dalzell, N. (2019). Classifying hate speech using a two-layer model. Statistics and Public Policy, 6(1), 80-86.

[5] Djuric N., Zhou J., Morris R., Grbovic M., Radosavljevic V., Bhamidipati N. Hate speech detection with comment embeddings; Proceedings of the 24th International Conference on World Wide Web; Florence, Italy. 18–22 May 2015; pp. 29–30.

[6] Mustafa, Raza & Nawaz, M. Saqib & Lali, Muhammad Ikram & Zia, Tehseen. (2017). Predicting The Cricket Match Outcome Using Crowd Opinions On Social Networks: A Comparative Study Of Machine Learning Methods. Malaysian Journal of Computer Science. 30. 10.22452/mjcs.vol30no1.5.

[7] Silva L., Mondal M., Correa D., Benevenuto F., Weber I. Analyzing the targets of hate in online social media; Proceedings of the Tenth International AAAI Conference on Web and Social Media; Cologne, Germany. 17–20 May 2016.

[8] Waseem Z., Hovy D. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter; Proceedings of the NAACL Student Research Workshop; San Diego, CA, USA. 13–15 June 2016; pp. 88–93.

[9] Badjatiya P., Gupta S., Gupta M., Varma V. Deep learning for hate speech detection in tweets; Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee; Perth, Australia. 3–7 April 2017; pp. 759–760.

[10] Davidson T., Warmsley D., Macy M., Weber I. Automated hate speech detection and the problem of offensive language; Proceedings of the Eleventh International AAAI Conference on Web and Social Media; Montreal, QC, Canada. 15–18 May 2017.

[11] Gambäck B., Sikdar U.K. Using convolutional neural networks to classify hate-speech; Proceedings of the First Workshop on Abusive Language Online; Vancouver, BC, Canada. 30 July-4 August 2017; pp. 85–90.

[12] Salminen J., Almerekhi H., Milenković M., Jung S.G., An J., Kwak H., Jansen B.J. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media; Proceedings of the Twelfth International AAAI Conference on Web and Social Media; Palo Alto, CA, USA. 25–28 June 2018

[13] Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, Pedro Henriques Hate Speech Classification in Social Media Using Emotional Analysis; Brazilian Conference on Intelligent Systems (BRACIS), 2018
[14] Yiwen Tang and Nicole Dalzell, Classifying Hate Speech Using a Two-Layer Model, Statistics and Public Policy, Volume 6, Issue 1,2019