A Comparison of Machine Learning Algorithms on Twitter Sentiment Analysis

Emon Ghosh	**180104134**
Tauhidur Rahman	**180104148**
Mohammad Arman	**180104151**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

A Comparison of Machine Learning Algorithms on Twitter Sentiment Analysis

Submitted by

Emon Ghosh	**180104134**
Tauhidur Rahman	**180104148**
Mohammad Arman	**180104151**

Submitted To

Faisal Muhammad Shah

Sajib Kumar Saha Joy,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

ABSTRACT

There are over 396.5 million Twitter users worldwide. In recent years racist, sexist, harassing tweets have been increased drastically. The number of abusers on Twitter are on the rise. So there's dire need of identifying the abusive tweets from the Twitter as it will help the new generation to be more careful about choosing their words more wisely. We used the Twitter Sentiment Analysis dataset [1] for our project collected from the Kaggle website. This dataset contains information about racist/sexist tweets. We have evaluated different machine learning algorithms such as Logistic Regression, Decsion Tree, Random Forest, KNN, SVM, Naive Bayes and Adaboost in this project and the results show that SVM algorithm perform well in the dataset.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The automated method of finding and categorizing subjective information in text data is known as sentiment analysis. This could be a viewpoint, a conclusion, or an emotion on a certain subject or product feature. The classification of emotions within a textual material is also known as opinion mining or emotion extraction. This method has been utilized frequently over the years to identify the attitudes and emotions contained in a specific textual material. The most common type of sentiment analysis is 'polarity detection' and involves classifying statements as Positive, Negative. Twitter is a social networking site that users typically use to express their feelings over specific events. With more than 396.5 million users and over 500 million tweets each day, Twitter has become a significant microblogging platform. With such a huge audience, Twitter continuously draws users who want to express their thoughts and perspectives on any problem, product, business, or other topic of interest. This is why a lot of businesses, institutions, and organizations use Twitter as a source of information. Users are able to express their opinions on Twitter using 140-character tweets. As a result, people tend to make their statements shorter by employing slang, acronyms, emoticons, short forms, etc. Along with this, people also use polysemy and sarcasm to express their thoughts.

Chapter 2

Literature Reviews

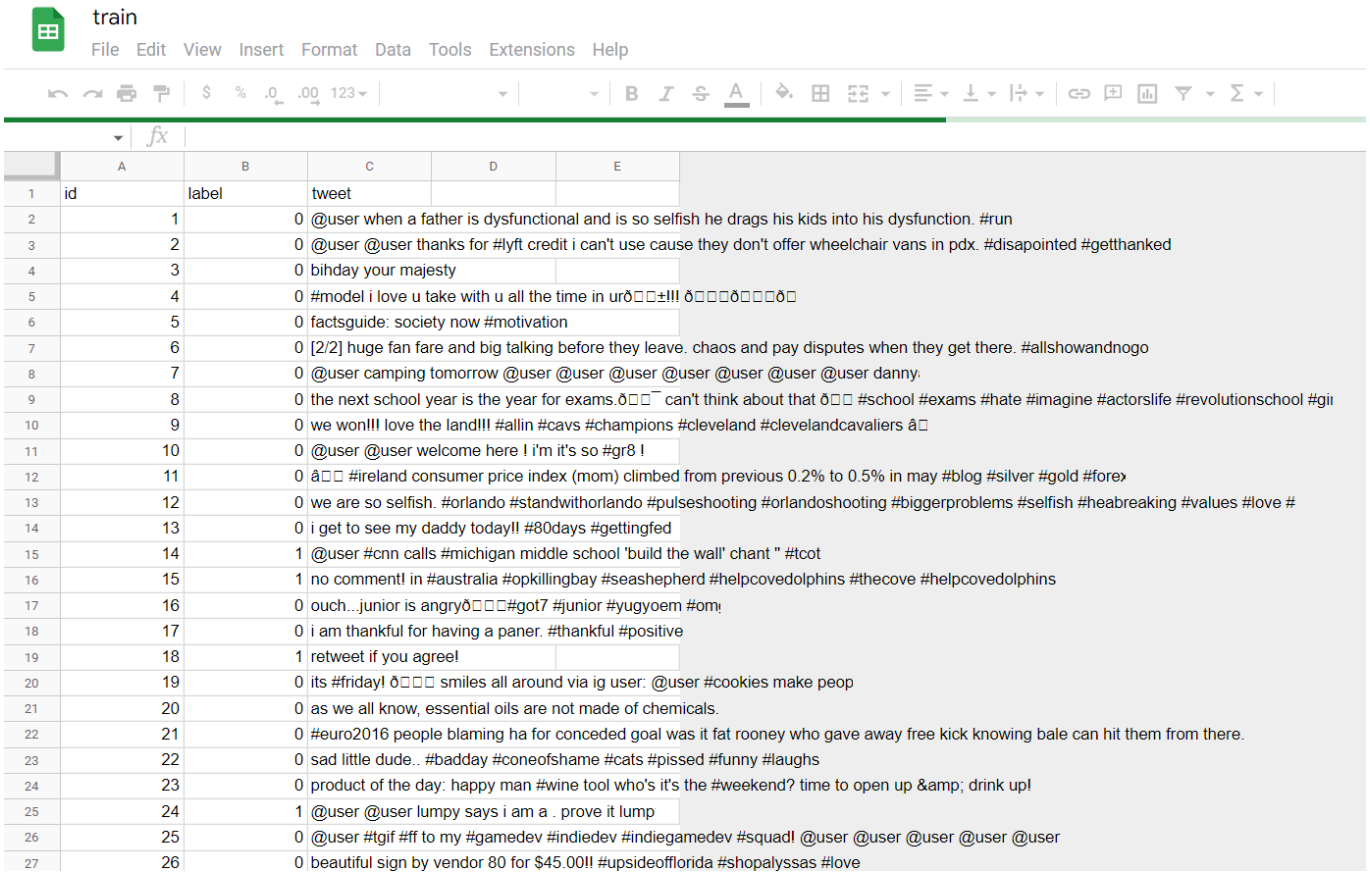
Several scholars used the machine learning (ML) method to sentiment analysis using different twitter sentiment dataset. Some closely related works are discussed in this section.
Gupta, Moniak Negi. et al. [2] showed 85.0% accuracy by applying the svm techniques and 66.24% by applying naive bayes classifier on their dataset.
Bac Le.et al[3] used showed 79.54% accuracy using naive bayes and 79.58% accuracy using svm on on standford sentiment dataset.
Geetika et al[4] used showed 88.024% accuracy using naive bayes and 85.55% accuracy using svm on amazon product review dataset.
In this work, to predict diabetes in a patient, different machine learning classification algorithms like Logistic Regression, SVM, K Nearest Neighbor (KNN) and Decision Tree (DT) are used and evaluated on the dataset. The evaluation of the performance of all the classification methods is done with various measurement methods.

Chapter 3

Data Collection & Processing

In our project, the twitter sentiment dataset [1] is collected from the Kaggle, which is originated from the post of twitter. In the sentiment dataset, the tweet contains hate speech if it has a racist or sexist sentiment associated with it. The dataset contains information about 31,962 tweets and their corresponding unique attributes. A training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist.

Preprocessing helps transform data so that a better machine learning model can be built, providing higher accuracy. The preprocessing performs various functions: outlier rejection, filling missing values, data normalization, feature selection to improve the quality of data. In the dataset, 29270 samples are classified as positive or not racist/sexist, and 2242 are classified as negative or racist/sexist.



train

File Edit View Insert Format Data Tools Extensions Help

	A	B	C	D	E
1	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 urð□□±!!! ð□□□ð□□□ð□		
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 @user danny.		
9	8	0	the next school year is the year for exams.ð□□¯ can't think about that ð□□ #school #exams #hate #imagine #actorslife #revolutionschool #gi		
10	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â□		
11	10	0	@user @user welcome here ! i'm it's so #gr8 !		
12	11	0	â□□ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #fore›		
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 angryð□□□#got7 #junior #yugyoem #om		
18	17	0	i am thankful for having a paner. #thankful #positive		
19	18	1	retweet if you agree!		
20	19	0	its #friday! ð□□□ smiles all around via ig user: @user #cookies make peop		
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 lump		
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		

Figure 3.1: Dataset Before Preprocessing

3.1 Dataset Preprocessing

Once the data is collected in the .csv files, it is very important to pre-process the information and remove all the unnecessary content. A number of pre-processing steps are involved, which are as follows:

1)Remove Pattern: This involves removal of urls, hashtags, numbers, special characters and punctuation from tweets. It is mainly done in order to clean the strings from unnecessary pattern.

2)Stemming: Here words are replaced by their stems or roots i.e. reading is replaced by read.

3) Tokenization: This involves removal of URLs, at-mentions and breaking of a string into a list of tokens. It is mainly done in order to count the occurrence of words. 4) Stop words removal: It involves removal of prepositions, articles that have high occurrence but do not have any influence on the overall sentiment of the text.

	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	when father dysfunct selfish drag into dysfunc...
1	2	0	@user @user thanks for #lyft credit i can't us...	thank #lyft credit caus they offer wheelchair ...
2	3	0	bihday your majesty	bihday your majesti
3	4	0	#model i love u take with u all the time in ...	#model love take with time
4	5	0	factsguide: society now #motivation	factsguid societi #motiv
5	6	0	[2/2] huge fan fare and big talking before the...	huge fare talk befor they leav chao disput whe...
6	7	0	@user camping tomorrow @user @user @user @use...	camp tomorrow danni
7	8	0	the next school year is the year for exams.ð□□...	next school year year exam think about that #s...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	love land #allin #cav #champion #cleveland #cl...
9	10	0	@user @user welcome here ! i'm it's so #gr...	welcom here
10	11	0	â□□ #ireland consumer price index (mom) climb...	#ireland consum price index climb from previou...
11	12	0	we are so selfish. #orlando #standwithorlando ...	selfish #orlando #standwithorlando #pulseshoot...
12	13	0	i get to see my daddy today!! #80days #getti...	daddi today #gettingf
13	14	1	@user #cnn calls #michigan middle school 'buil...	#cnn call #michigan middl school build wall ch...
14	15	1	no comment! in #australia #opkillingbay #se...	comment #australia #opkillingbay #seashepherd ...

Figure 3.2: Preprocessed Dataset

3.2 Positive and Negative words in Twitter dataset

We have extracted some postive words such as love, life, smile, friend etc from the dataset. We have also plot the frequently used positive words in the dataset.

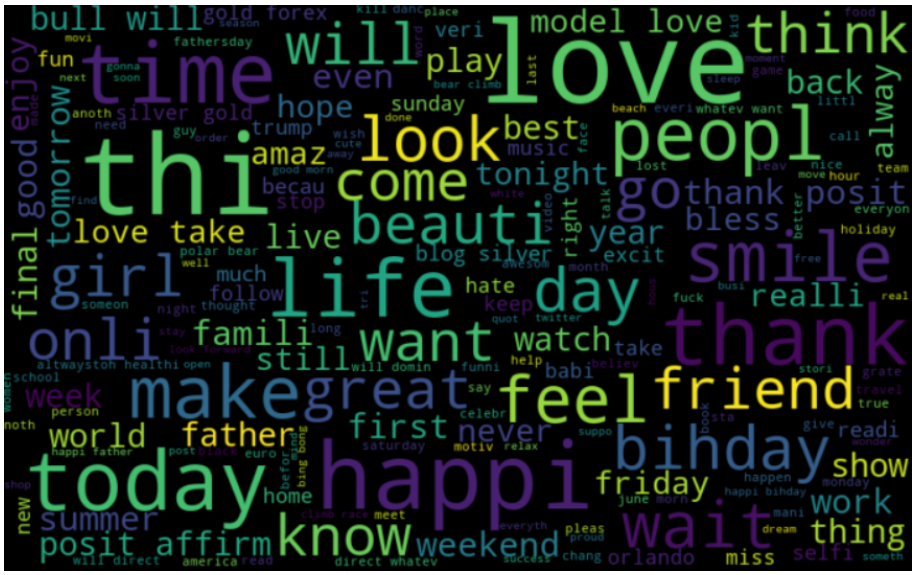


Figure 3.3: Positive words in the dataset

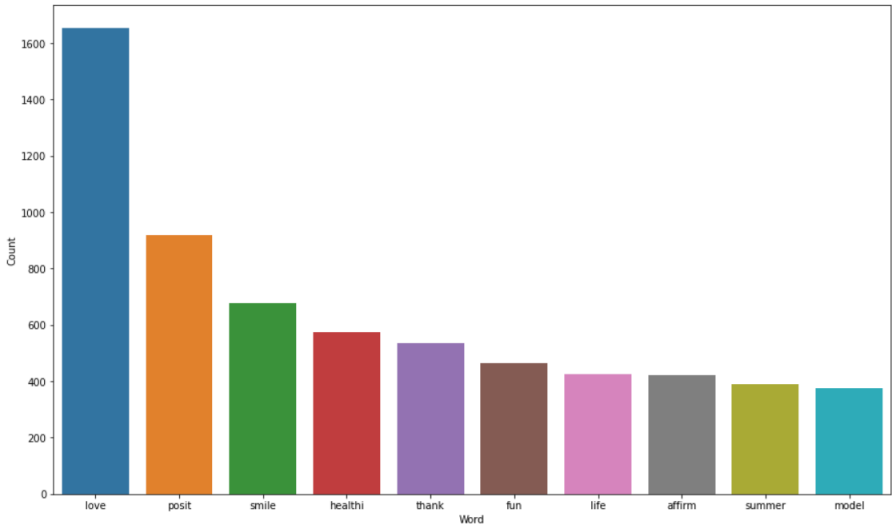


Figure 3.4: Positive word Frequency

We have extracted some negative words such as Trump, libtard, racist, fuck etc from the dataset. We have also plot the frequently used negative words in the dataset.

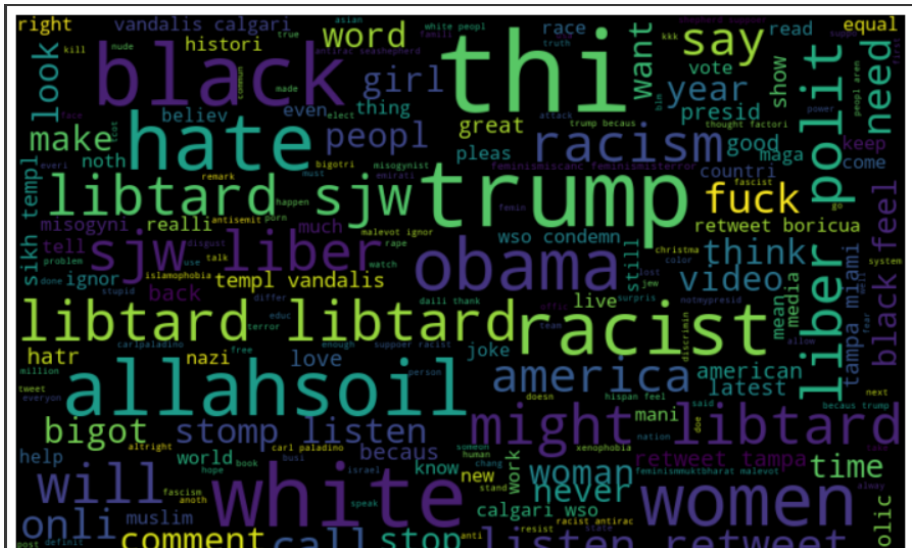


Figure 3.5: Negative words in the dataset

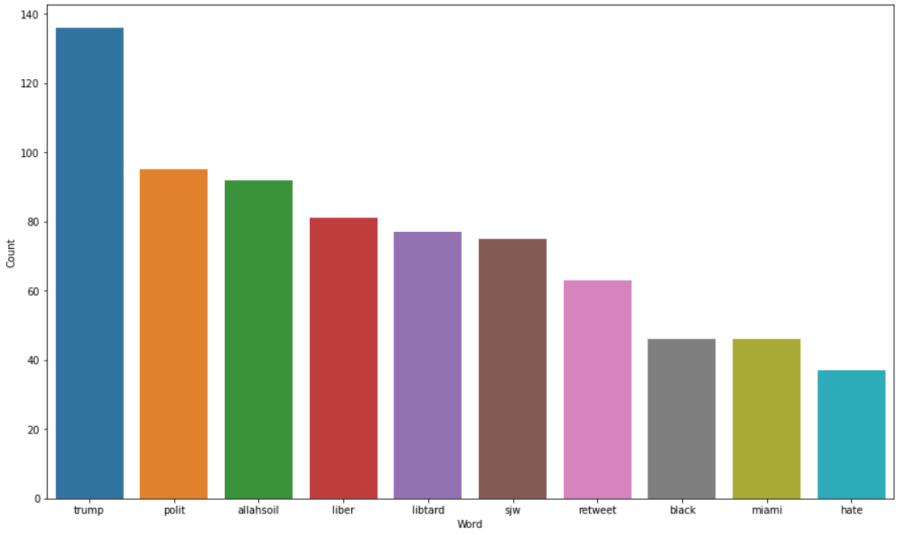


Figure 3.6: Negative word Frequency

Chapter 4

Methodology

4.1 Proposed Model

For our project we used four machine learning algorithms like Logistic Regression, Support Vector Machine (SVM), K Nearest Neighbor(KNN), Decision Tree(DT), Random Forest (RF), Naive Bayes and AdaBoost. For this we preprocessed the data and split the data for training and testing the model. For training we used 70% of the data and rest of the 25% for testing purpose. After training the models, we evaluated the models with testing data.

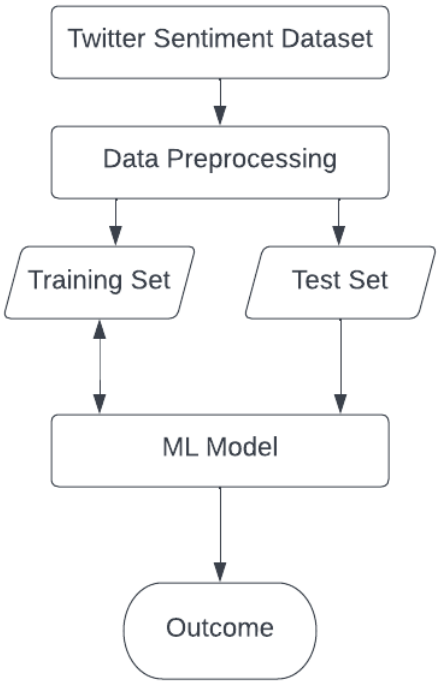


Figure 4.1: Methodology

4.2 Train Test Split

We have used 75% data to train our model and 25% data to test our model.

4.3 Feature Extraction

When modeling text using machine learning techniques, one method of encoding text data is known as the bag-of-words model. It is a Feature Extraction method of converting the text data into numerical vectors as features.The method is really straightforward and adaptable, and it may be applied in a variety of ways to extract features from documents.

Chapter 5

Experiments and Results

5.1 Random Forest Classifier

Random Forest is a supervised learning algorithm that can be used for both classification and regression problems. It is a three structure classifier that 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 model gave an accuracy of 94.20% on the training data.

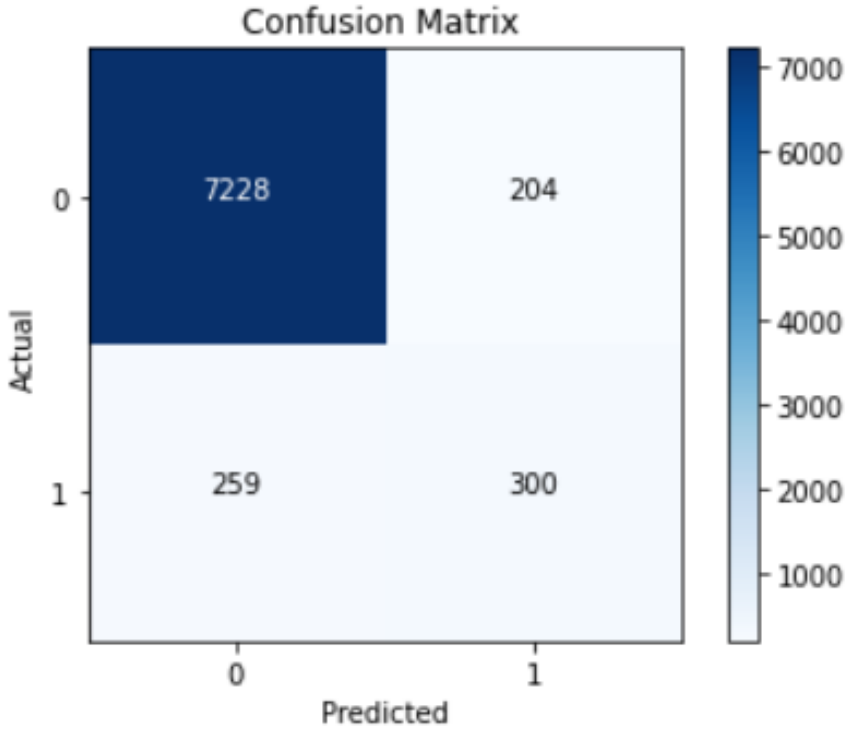


Figure 5.1: Confusion Matrix of Random Forest

5.2 Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

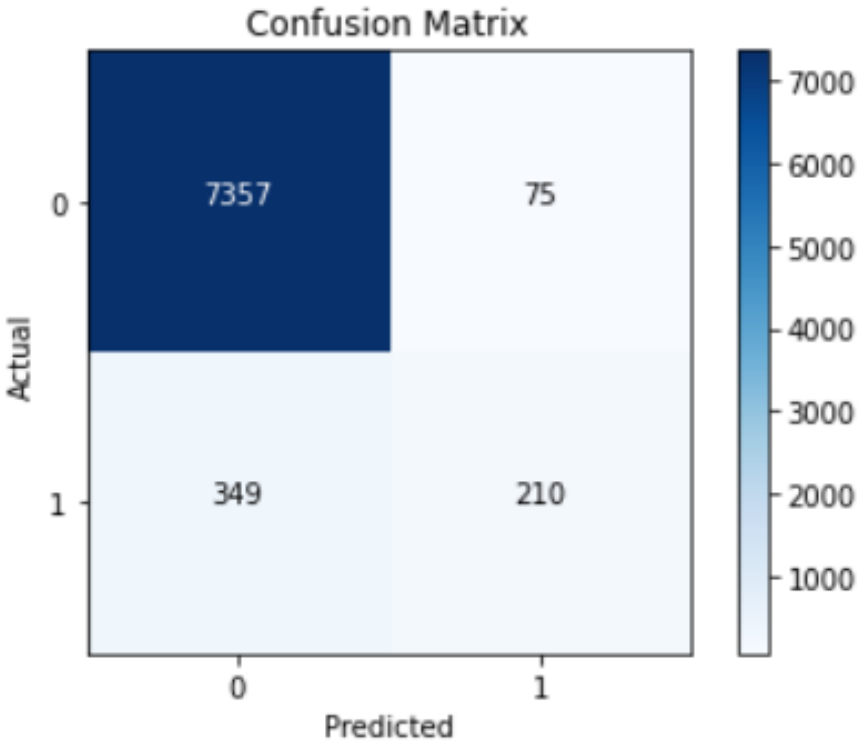


Figure 5.2: Confusion Matrix of Logistic Regression

Logistic Regression achieved 94.69% accuracy on the training data.

5.3 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. We used this algorithm to predict the tweets. The model achieved an accuracy 94.73 % accuracy on the training data.

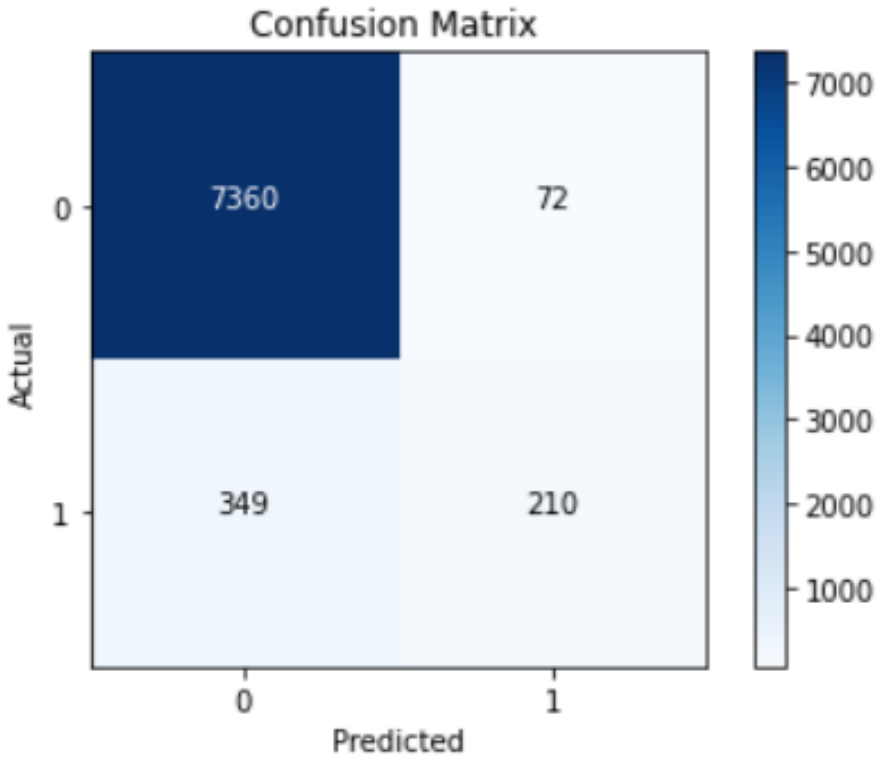


Figure 5.3: Confusion Matrix of SVM

5.4 K Nearest Neighbours

We have used the K-Nearest Neighbor classifier which is also known as instance-based classifier and belongs to a family of learning methods called lazy learning. In K-nearest neighbor (KNN), the classifier evaluates distance from the test data point (also known as the query data point) to all the data points in the training set. Following this, it finds the K nearest neighbors to this test data point. A simple voting is then conducted between the K nearest data points to decide on the class that the classifier predicts. The model gave an accuracy of 94.05% accuracy on the training data.

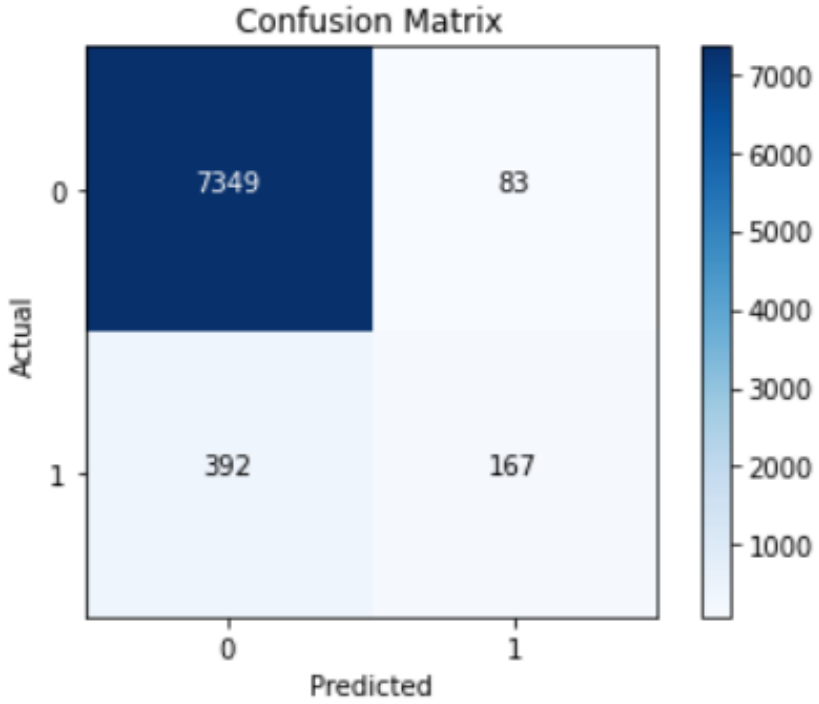


Figure 5.4: Confusion Matrix of KNN

5.5 Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification 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. This model gives an accuracy of 93.14 % on the training data.

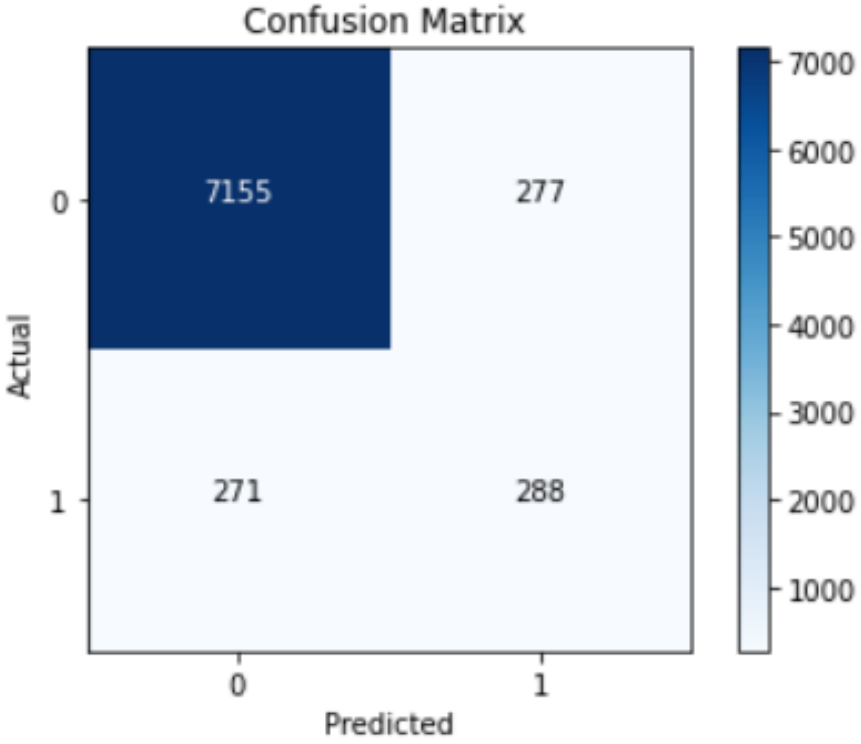


Figure 5.5: Confusion Matrix of Decision Tree

5.6 Naive Bayes Classifier

This is one of the supervised algorithm that is based on the probabilistic approach to classify the text to a particular class i.e. positive or negative. This algorithm calculates the probability of all the words in the dataset and then classifies the tweets or text into particular categories.This model gives an accuracy of 93.56 % on the training data.

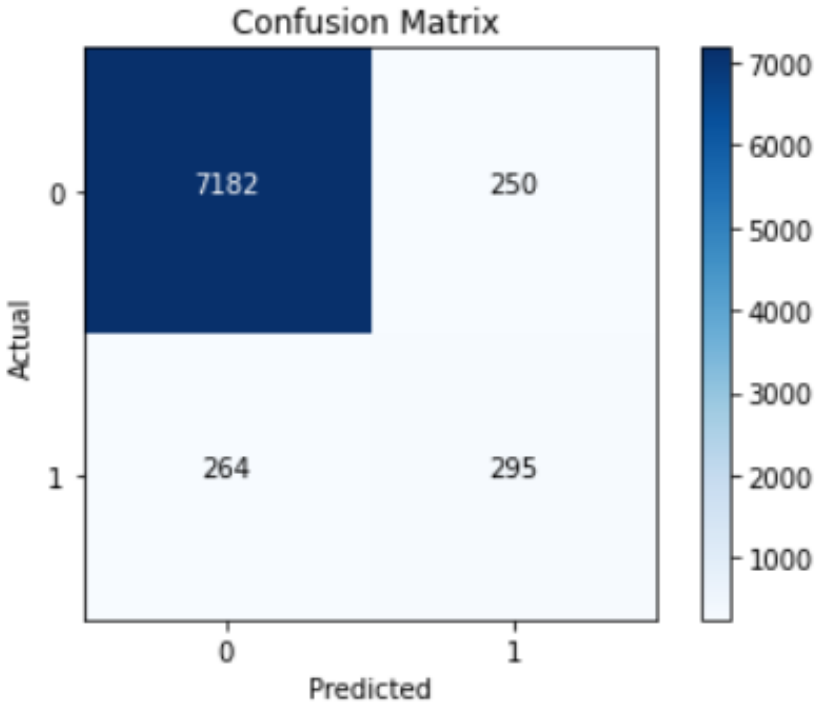


Figure 5.6: Confusion Matrix of Naive Bayes Classifier

5.7 Adaboost Classifier

The Boosting technique known as AdaBoost algorithm, sometimes known as Adaptive Boosting, is used as an Ensemble Method in machine learning. The weights are redistributed to each instance, with higher weights being given to instances that were mistakenly categorised, hence the name "adaptive boosting." For supervised learning, boosting is used to lower bias and variation. This model gives an accuracy of 94.44 % on the training data.

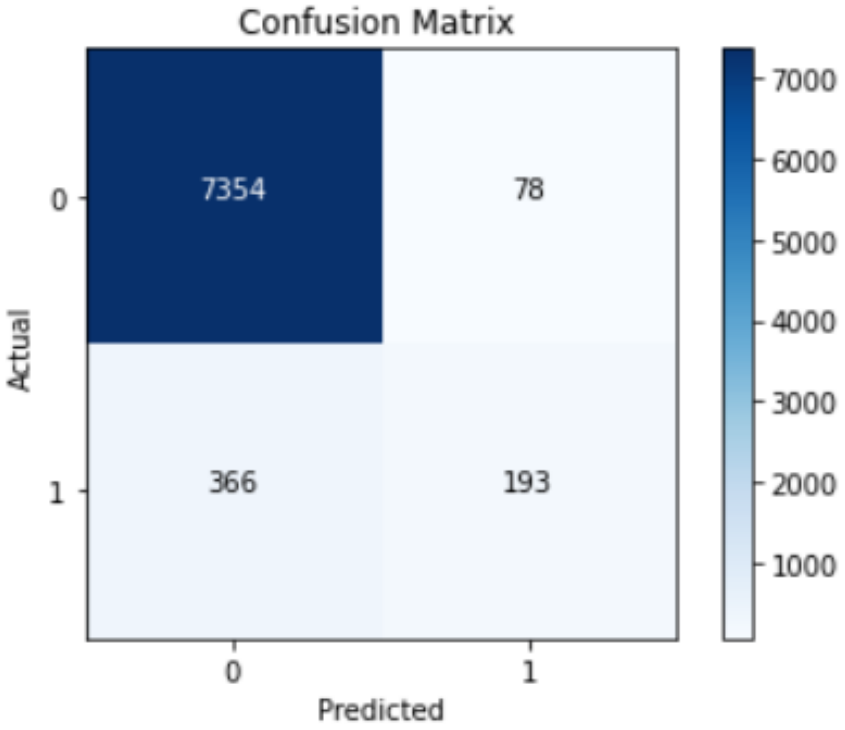


Figure 5.7: Confusion Matrix of Adaboost Classifier

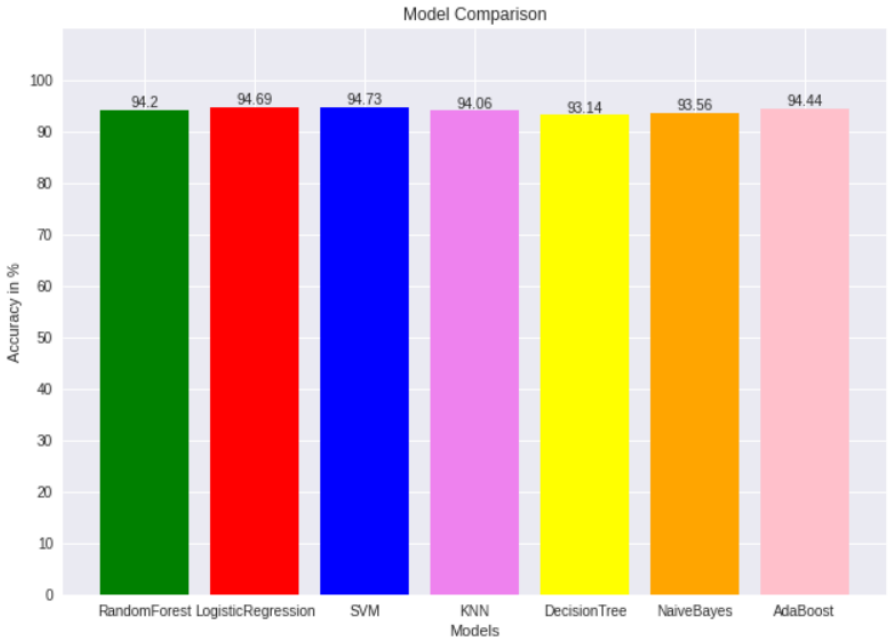


Figure 5.8: Accuracy of each model on Test Dataset

From the Fig:5.8, we can see Random Forest, Logistic Regression, SVM, KNN, Decision Tree, Naive Bayes and Adaboost models have corresponding accuracy of 94.2% , 94.69% , 94.73% , 94.06% , 93.14% , 93.56% , 94.44% . We can see that SVM performs better than all other models on classifying racist/sexist tweets from Twitter Sentiment Dataset.

5.8 ROC Curve

The receiver operating characteristic curve (ROC curve) is a graph that displays how well a classification model performs across all categorization levels. True Positive Rate and False Positive Rate are the two parameters plotted on this curve. It demonstrates the trade-off between specificity and sensitivity (or TPR) (1 – FPR). A better performance is shown by classifiers that provide curves that are closer to the top-left corner.

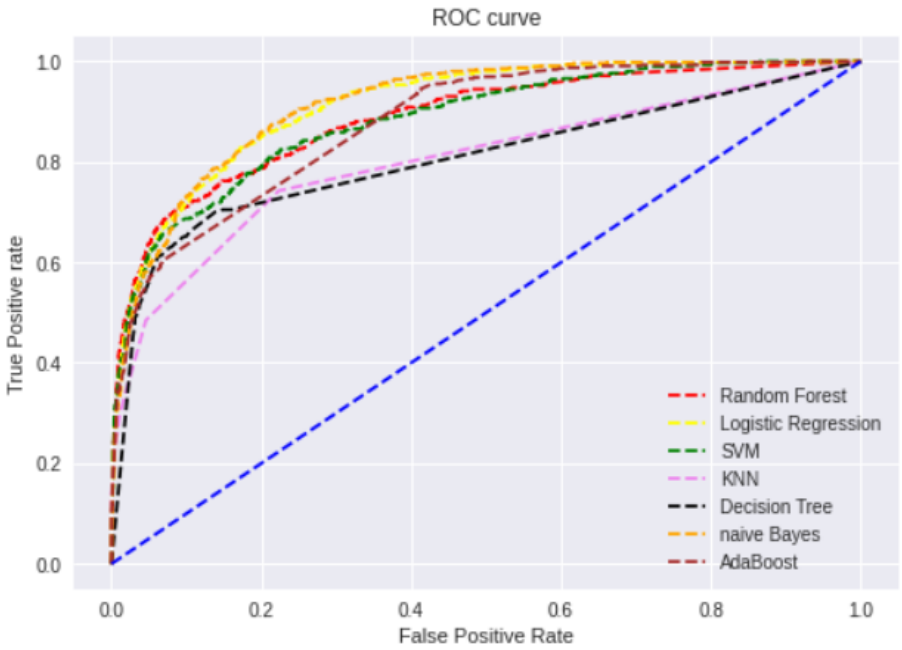


Figure 5.9: Receiver Operating Characteristic(ROC) Curve

Chapter 6

Future Work and Conclusion

In this project we have used the twitter sentiment dataset and experimented with seven machine learning algorithms like Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, K Nearest Neighbours (KNN), Random Forest Classifier, Naive Bayes and AdaBoost Classifier. We trained and evaluated the stated model and found out SVM model performed better than other models on classifying racist/sexist tweets from the twitter sentiment dataset. The SVN model achieved an accuracy of 94.73

In future, we plan to collect a more enriched dataset which will help us to classify racist/-sexist tweets with higher confidence. In addition to this, we also plan to extend this work to evaluate how complex classifiers based on Artificial Neural Network (ANN) or other deep learning techniques perform.

References

[1]https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech
[2]Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python, Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani
[3]Twitter Sentiment Analysis Using Machine Learning Techniques, Bac Le and Huy Nguyen
[4]Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis, Geetika Gautam, Divakar yadav

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Wednesday 31st August, 2022 at 3:47am.