Twitter Dataset Based Sentimental Analysis with Machine Learning Approach

Endluru Vignesh
AIT-CSE, AIML
Chandigarh University, Punjab, India
endluruvignesh2002@gmail.com

Kethuri Ajay
AIT-CSE, AIML
Chandigarh University, Punjab, India
ajaykethuri@gmail.com

Vupputuri Bharath
AIT-CSE, AIML
Chandigarh University, Punjab, India
bharathupputuri@gmail.com

Koyyada Akshay Kumar
AIT-CSE, AIML
Chandigarh University, Punjab, India
akshay.k0520@gmail.com

Abstract- The aim of this study is to observe and pointing out the results based on it. This study shows how data pre-processing steps are working inside and how this Natural Language Processing (NLP) techniques are efficient for sentiment analysis. Here a model is created for sentiment analysis based on twitter data and NLP functions and techniques. Using profuse machine learning models such as Logistic Regression, Support Vector Machine, Multinomial Naïve Bayes, Random Forest, K Nearest Neighbor and Linear Discriminant Analysis tweets are classified to respective labels and then trained and tested. Analyzing the results and giving the conclusion based on that is key for this study.

I.INTRODUCTON

The reports say that, millions of people are using Social networks like Twitter, Instagram and Facebook websites to express their feelings, personal opinions and publish about their daily lives. However, people express or write different tweets on different platforms such as sharing any good news or some achieved things in their life. Sentiment Analysis which is done under Natural Language Processing that identifies the emotion of the text that a particular person posted.
Currently, the advent of Internet has transformed the way individuals communicate their thoughts, opinions. In the sentiment analysis it is very important to data pre-process because it enhances the efficiency and accuracies of the models used further. Generally, data pre-processing is used by different NLP techniques like removal of unnecessary words and tokenization is used for further greater understanding for the machine.

This project also analyzes the different machine learning models and classifiers on how they are working for the given data. Here classifiers used for classification are Logistic Regression, SVM, Multinomial Naïve bayes, Random Forest, K Nearest Neighbor and Linear Discriminant Analysis.

The primary objective of sentiment analysis is to extract meaningful insights from large volumes of unstructured text data. By automatically assessing the sentiment behind the text, sentiment analysis enables businesses and organizations to gain valuable insights into public opinion, customer satisfaction, and brand perception. This information can be utilized to make data-driven decisions, improve products or services, enhance customer experience, and identify emerging trends or issues.

II.LITERATURE SURVEY

Sentiment analysis is a major approach for classifying the given text into positive and negative in Natural Language Processing (NLP).

In [1] 2023, Zahratu Sabrina explained what are the different approaches for sentiment analysis. It includes Lexicon-based approach, which works by splitting the sentences into bag of words and compare them with the words of sentiment polarity lexicon and semantic relations. Second methos used is machine learning approach, which uses classifiers to extract features from the data. It includes data collecting, data pre-processing, extracting features, training data and analyzing results.

In [2] 2023, Shamsuddeen Hassan Muhammad implemented sentiment analysis for various languages in Africa. It explains the challenges for African languages in sentiment analysis. Here many local African languages which are very new to the sentiment analysis are the inputs for the model. This contributed majorly for Non-English language.

In [3] 2021, Vedurumudi Priyanka used ensembling technique, which combines various classifiers, in order to improve the accuracy for the dataset taken from Kaggle. Dataset consists of tweet_id, sentiment, and tweet where tweet_id is the unique number for every observation. Data is pre-processed followed by extracting features from the pre-processed data. Features are extracted using n-grams procedure. Features are represented using Sparse vector and Dense vector.

In [4] 2023 Imane Lasri used sentiment analysis for Moroccan public universities from twitter using big data technologies. Collected data from twelve public Moroccan universities and pre-processed data to label them. 1798 tweets were taken using Twint[32], an open python library to extract tweets based on keywords.

In [5] 2022, Yili Wang used a few other approaches for sentiment analysis which includes probabilistic classifier, Linear classifier, rule-based classifier, hybrid approach and other approaches. It concludes that machine learning approaches are more efficient than other approaches. The goal here is to survey on the existing methods for sentiment analysis and it is achieved.

In [6] 2022, Masoud AminiMotlagh collected data, pre- processed it, detected and classified accordingly. This research paper included some of the machine learningtechniques like KNN, SVM, NAÏVE BAYES and Bagging. Various test splits are used for each technique to check the accuracy and how the classifiers are working in many cases. At 70% train test split is working efficiently and for voting technique 85.71 accuracy is extracted, which is the highest among all cases. Variations are also taken out from the different cases. Results are represented usinga line graph.

In [7] 2022, Astha Modi analyzed how different techniques like LSTM, Naïve bayes, decision tree and SVM are working. Accuracies are also recorded and future scopes of every technique also mentioned in it. Data is analyzed based on scenarios. Tweets are taken on IPL, OTT service providers like Netflix, amazon prime etc, Footwear companies like Nike and Adidas, Indian industries reliance and adani. Results and analysis are represented in pie charts.

In [8] 2019, Abdul Rasool took a case study on twitter sentiment analysis for apparel brands. Firstly, data is pre-processed and cleaned. Two brands are taken namely Nike and Adidas. 54788 tweets are taken for Nike and 45062 tweets are taken for Adidas. Totally 99850 tweets are taken. In this research paper Naïve bayes classifier is used to analyse and obtain patterns from it. Results explains on the percentages of positive, negative and neutral from both brands.

In [9] 2019, Faizan followed a certain procedure which includes data collection, data pre-processing, feature extraction, model selection and model evaluation. API available for twitter is used to collect data from twitter. By using data pre-processing techniques, unimportant data like hashtags, @usernames is removed. Feature extraction is done by POS tagging and n-grams techniques. KNN technique is used for model selection where it classifies data into its target values. Finally, model is evaluated using confusion matrix.

In [10] 2019, Abdullah Alsaeedi also used classification techniques that are Naïve bayes, Maximum Entropy and Support Vector Machine. Differentiation of document level and sentence level sentiment analysis also addressed in this research paper. After training most of the data then, model is built. Remaining data is tested by labels and analysis is taken out from the results.

III. PROPOSED SYSTEM

Here in this study, we are trying to show how twitter data-based sentiment analysis is working using NLP techniques and different Machine Learning Models for classifying the extracted data. Hence, this drives us to create and analyze the models on this data. Public tweets from twitter are taken out by manually or using API. Collected data is pre-processed using many NLPtechniques like POS tagging, Stop word removal etc. After collecting pre-processed data, it is loaded into different machine learning models for classification. Lastly, observing results we can conclude how the models are working for tweets from twitter. This helps us to know the trends on some tweets and how people are responding for it.

IV. METHODOLOGY

In this study there are primarily 5 stages to it. Every stage is so vital and helps in building the efficient model. Following below are the steps:

A. Data Collection.

B. Data Preprocessing.

C. Feature Selection.

D. Model Selection.

E. Model Evaluation.

All the steps mentioned above are observed keenly and tried to make them more added assets for the study.

A. *Data Collection*

Data in this study is extracted using a twitter API where it is used for retrieving tweets on our required basis. A twitter developer account is created and requested for access to API. After request is accepted, we can use the twitter API and tweepy or twint libraries and functions are used to extract n number of tweets automatically and saved to a csv file.

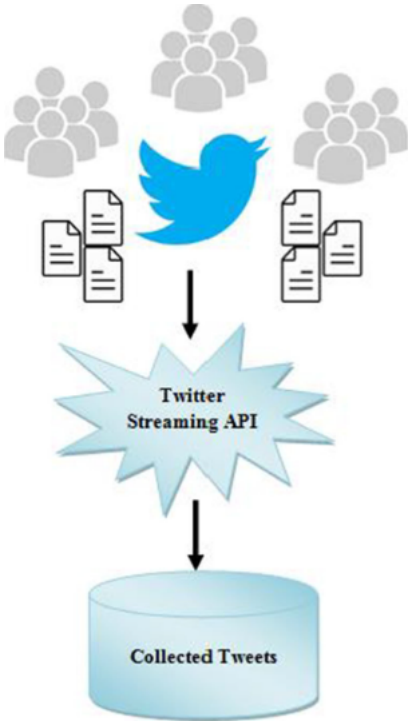

FIGURE 1: API PROCESS

Twitter API provides the unique ids for connecting the tweepy or twint library to our API account which includes consumer key, consumer secret key, access key and access secret key. If these tokens match then, we can access the tweets and retrieve tweets from twitter.

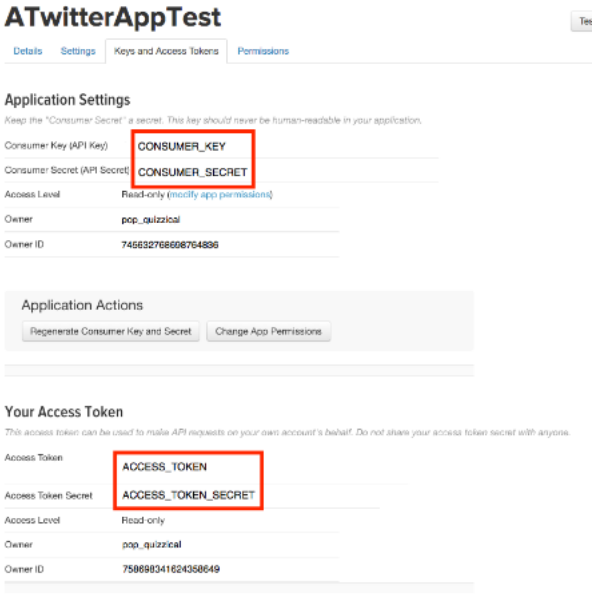

FIGURE 2: TOKENS AND KEYS API

B. *Data Preprocessing*

The preprocessing is an essential process where raw data taken from twitter is processed and presented into more convenient format. This preprocessed data is loaded into classifier later. There is always a scope for filtering unwanted data or the data which restricts our model to perform well. There are some steps in this data preprocessing method like the following:

- Hashtags: these are the common words in twitter followed by "#". These are the unnecessary words of data where the model cannot process and are unwanted. So, these hashtags are removed from data using a vectorize function where the specified character is removed from the data.
- @username: @ is another unwanted character which is used to mention the account holder name. This is also an unwanted and restricts our model to perform well. So, this also removed using same vectorize function.

	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 sel...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause th...
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 ...
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation

FIGURE 3: @ REMOVAL

- Characters: some unnecessary characters like ""?$ kind of punctuations are also removed using a replace function where those are replaced by a space.
- There are other few words basically which are having length less than 3 like a, an, the, like words which are insignificant those are also replaced by a space.

	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	when father dysfunctional selfish drags kids i...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks #lyft credit cause they offer wheelchai...
2	3	0	bihday your majesty	bihday your majesty
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	factsguide society #motivation

FIGURE 4: CHARACTER REMOVAL

- Emoticons: these are the symbols which are used to express the facial expressions using punctuations and other symbols. These does not have a minor role also in the data to be used for this study. These are also removed.

After these processes the significant words are left and this data is ready to be processed further.

C. *Feature Selection*

In this process some features are extracted and be used for model selection. Here few Natural Language Processing techniques are used to extract the features of the model.

Frist step is that every word is now differentiated

and represented in single words separated by commas. Word tokenization technique is used to tokenize the words from the data.

```
0    [when, father, dysfunctional, selfish, drags, ...
1    [thanks, #lyft, credit, cause, they, offer, wh...
2                              [bihday, your, majesty]
3                     [#model, love, take, with, time]
4                   [factsguide, society, #motivation]
```

FIGURE 5: TOKENIZATION

You can see in the above figure that each word is separated by commas for further process.

Now a stemming process is being done by using a stemmer from NLTK library where all techniques and functions are available for NLP. Stemming process is where it reduces the word to its root word called stem. In this process all prefixes, suffixes and infixes are removed and reduced to its root word. It helps in specifying the exact word which are crucial for further process.

```
0    [when, father, dysfunct, selfish, drag, kid, i...
1    [thank, #lyft, credit, caus, they, offer, whee...
2                              [bihday, your, majesti]
3                     [#model, love, take, with, time]
4                         [factsguid, societi, #motiv]
Name: clean_tweet, dtype: object
```

FIGURE 6: STEMMING

As you can see in the above diagram all affixes are removed and only significant words are the output. Lancaster stemmer is used here for stemming process.

There are other features used like unigram, bigram, n-gram, POS tagging are also used for further extraction on features. In these processes conditional probability is used to count the significance of the words.

D. Model Selection

Here different machine learning models are used. they are as follows:

1) Logistic Regression
2) Support Vector Machine
3) Multinomial Naïve Bayes
4) Random Forest
5) K Nearest Neighbor
6) Linear Discriminant Analysis

1). In this project, the first model used for the outcome of the project is the Logistic regression.

Logistic regression is used to classify the labels from the given data. Logistic regression is basically Boolean type model where it says the input belongs to label 1 or label 2.

The extracted features are loaded into this model and it is being processed and output is given.

First using train_test_split function our data is split into two parts which are used for training the data and testing the data. It is split into 80:20 percent because it is the ideal split for logistic regression.

Using the features logistic regression is trained by 80 percent of the data and it stores the features like specific words for a label. Further in testing it uses those features which are the main words for that label. It classifies the text given by searching those words and if it finds it classifies to respective label or sentiment.

2). Next used is SVM classifier which uses the kernel trick. Kernel trick finds the optimal line of the entire data. It classifies the data by creating a hyperplane and labels are opposite side of the hyperplane. If a data point is given for testing it decides which side it should lie.

3). Multinomial Naïve Bayes is very efficient for sentiment analysis since it uses bayes theorem in return it uses conditional probability.

4). Random Forest is a powerful classifier because it uses decision trees and ensemble learning. It uses numerous classifiers to classify the data. It splits the data randomly and averages the inside models.

5). K Nearest Neighbor classifier also used to analyze how it is behaving for the given data since it is also widely used classifier. It uses proximity for classification.

6). Linear Discriminant Analysis is a recent classifier. It generally used for dimensionality reduction and as a classifier. It is used for versatility of the project.

E. Model Evaluation

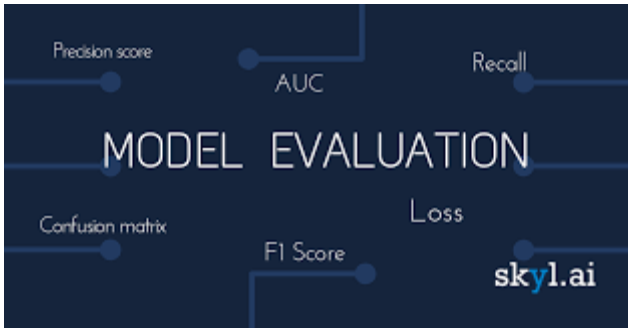

FIGURE 7: MODEL EVALUATION

Model evaluation is the final stage of the project. This shows the results by calculating some of the main parameters for a model. It calculates the accuracy, confusion matrix, precision score, recall, F1 score and so on. The key parameters are accuracy and confusion matrix. Both majorly signifies how the model is working first confusion matrix is calculated using formulas. Using the information from confusion matrix accuracy is measured. Accuracy says how well the model is classifying the input tweets or the text.

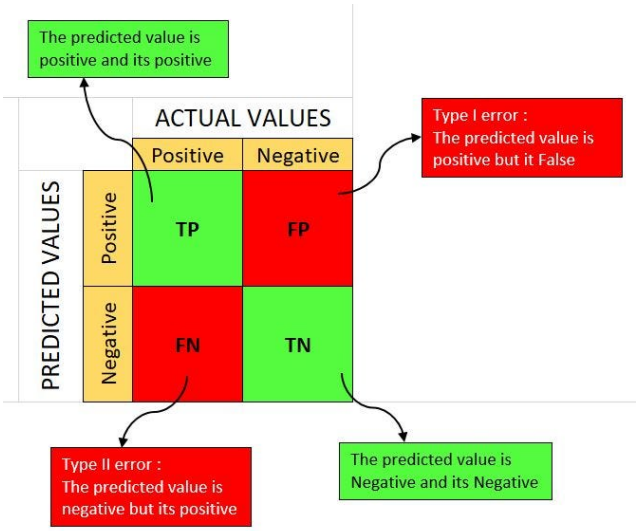

FIGURE 8: CONFUSION MATRIX STRUCTURE

By this step everything is completed and we are ready to analyze the results and conclude based on it.

V. RESULTS/OUTPUTS

The project and study created by us showed a significant accuracy of 94.4 percentage for logistic regression.

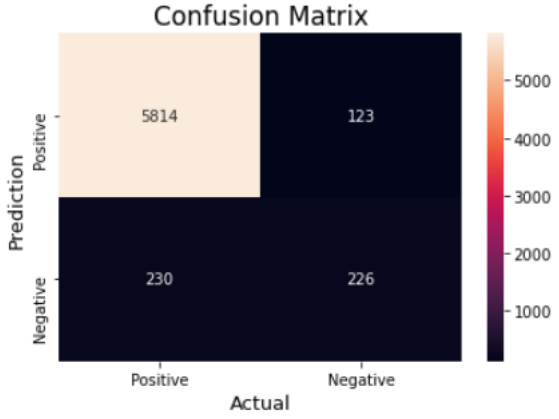

FIGURE 9: CONFUSION MATRIX OF LOGISTIC REGRESSION

This is the confusion matrix structure and it shows the true positive, true negative, false positive and false negative for our model.

TP= 5814

FP= 123

FN= 230

TN= 226

By using the above data, we calculate F1 score, precision score and recall.

Precision = TP/(TP+FP)

= 5814/5937

= 0.979

Recall = TP/(TP+FN)

= 5814/6044

= 0.961

F1 score = 2*precision*recall/(precision+recall)

= 2*0.979*0.961/0.979+0.961

= 0.5614

We represented a wordcloud for both positive and negative. Wordclouds for both are below.

FIGURE 10: WORDCLOUD FOR POSITIVE

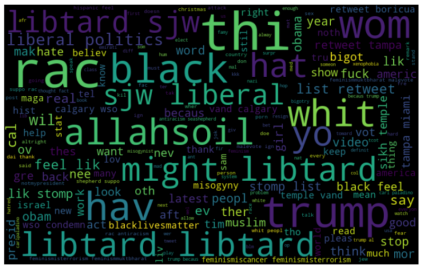

FIGURE 11: WORDCLOUD FOR NEGATIVE

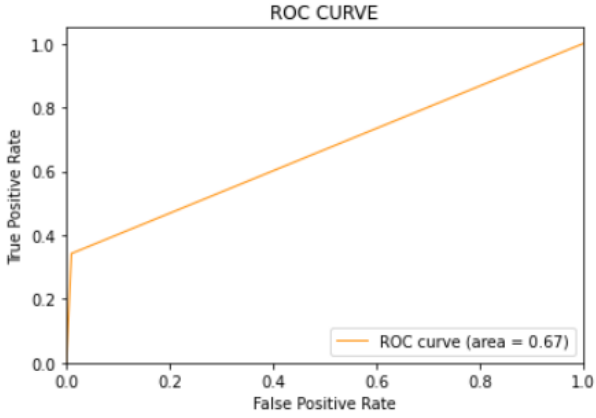

FIGURE 12: ROC LOGISTIC REGRESSION

SVM classifier achieved the accuracy of 94.50 percentage.

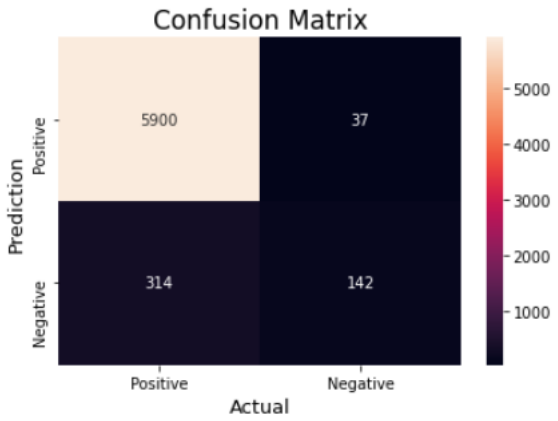

FIGURE 13: SVM CONFUSION MATRIX

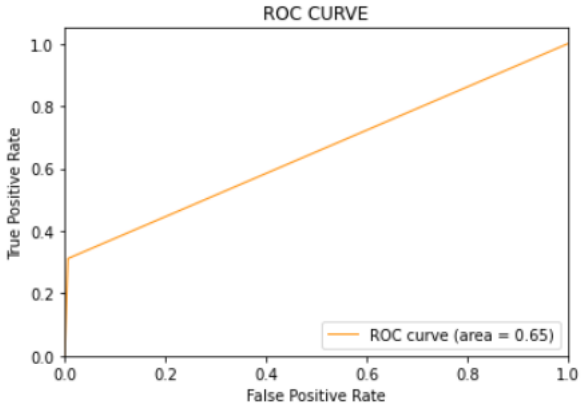

FIGURE 14: ROC SVM

Multinomial Naïve Bayes achieved the accuracy of 93.69 percentage.

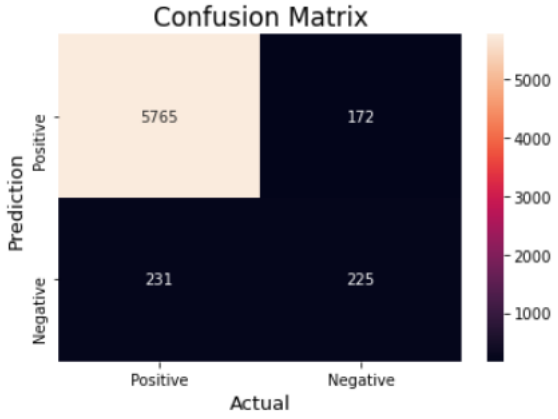

FIGURE 15: MULTINOMIAL NAÏVE BAYES CONFUSION MATRIX

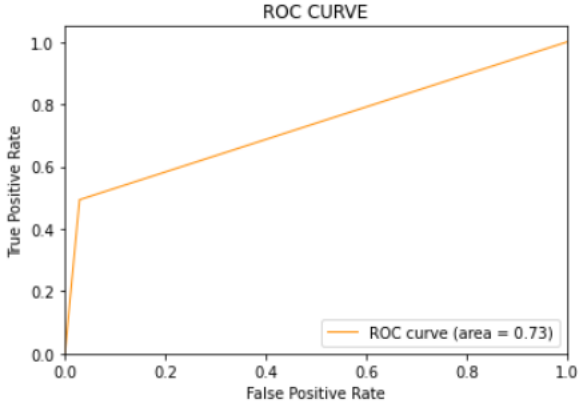

FIGURE 16: ROC MULTINOMIAL NAÏVE BAYES

Random Forest achieved the accuracy of 94.02 percentage.

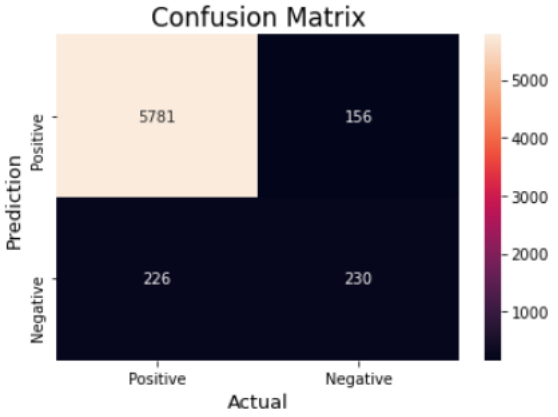

FIGURE 17: RANDOM FOREST CONFUSION MATRIX

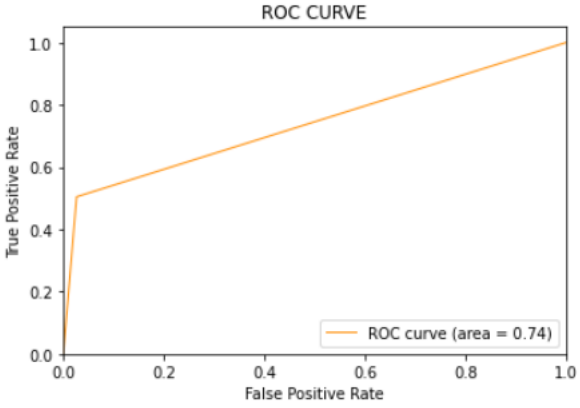

FIGURE 18: ROC RANDOM FOREST

K Nearest Neighbor achieved the accuracy of 93.79 percentage.

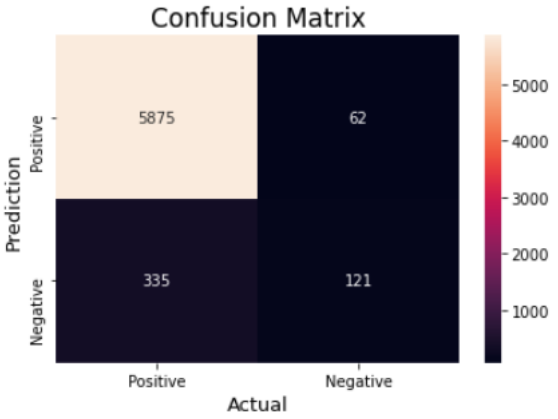

FIGURE 19: K NEAREST NEIGHBOR CONFUSION MATRIX

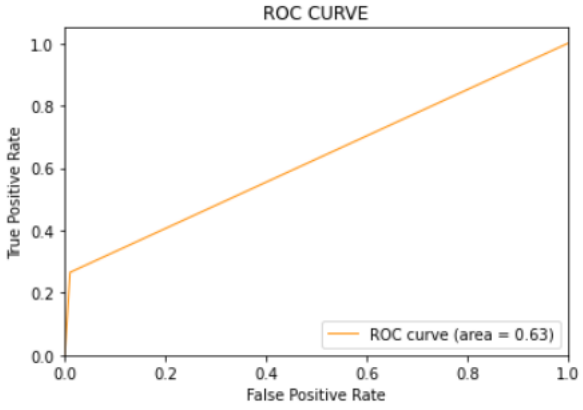

FIGURE 20: ROC K NEAREST NEIGHBOR

Linear Discriminant Analysis achieved the accuracy of 93.99 percentage.

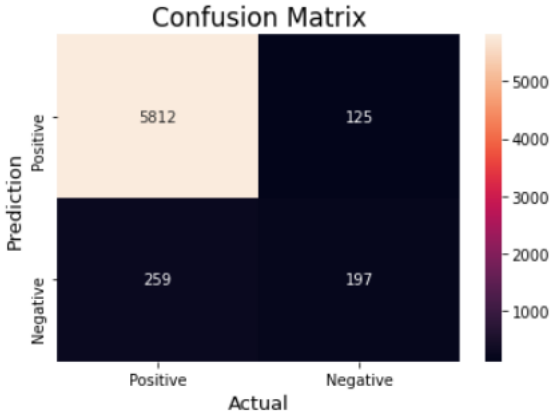

FIGURE 21: LDA CONFUSION MATRIX

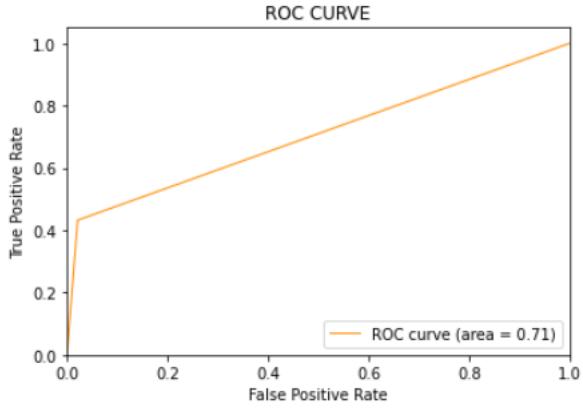

FIGURE 22: ROC LDA

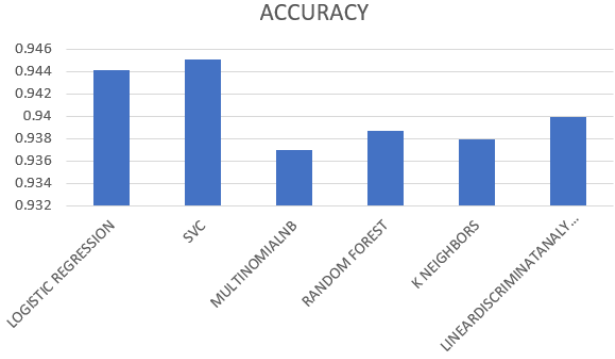

FIGURE 23: ACCURACY COMPARISON

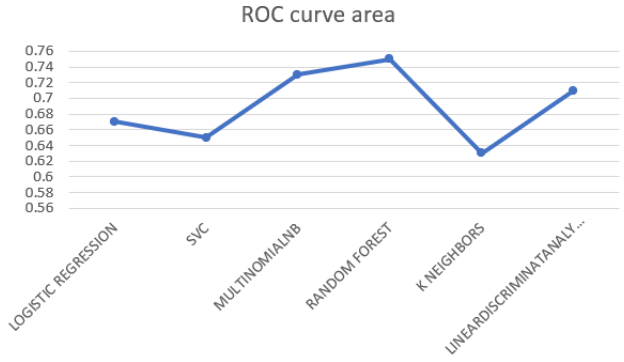

FIGURE 24: ROC CURVE COMPARSION

VI. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated how the twitter sentiment analysis is done using different machine learning models and Natural Language Processing (NLP). We performed training and testing on the data and achieved a good accuracy of 94 percentage. In the future this can be used or modified or enhanced for the purpose of extracting the twitter trends on a particular topic. With this model we can know how twitter people are reacting on a specific topic. This can be also helpful for further other classification techniques or deep learning or neural networks.

VII. REFERENCES

[1]. Zahratu Sabrina explained what are the different approaches for sentiment analysis, 2023.

[2]. 2023, Shamsuddeen Hassan Muhammad implemented sentiment analysis for various languages in Africa.

[3]. 2023, Vedurumudi Priyanka used ensembling technique, which combines various classifiers, in order to improve the accuracy for the dataset taken from Kaggle.

[4]. 2023 Imane Lasri used sentiment analysis for Moroccan public universities from twitter using big data technologies.

[5]. 2022, Yili Wang used a few other approaches for sentiment analysis which includes probabilistic classifier, Linear classifier, rule-based classifier, hybrid approach and other approaches.

[6]. 2022, Masoud AminiMotlagh collected data, pre-processed it, detected and classified accordingly. This research paper included some of the machine learningtechniques like KNN, SVM, NAÏVE BAYES

and Bagging.

[7]. 2022, Astha Modi analyzed how different techniques like LSTM, Naïve bayes, decision tree and SVM are working. Accuracies are also recorded and future scopes of every technique also mentioned in it.

[8]. 2019, Abdul Rasool took a case study on twitter sentiment analysis for apparel brands. Firstly, data is pre-processed and cleaned. Two brands are taken namely Nike and Adidas.

[9]. 2019, Faizan followed a certain procedure which includes data collection, data pre-processing, feature extraction, model selection and model evaluation. API available for twitter is used to collect data from twitter.

[10]. 2019, Abdullah Alsaeedi also used classification techniques that are Naïve bayes, Maximum Entropy and Support Vector Machine. Differentiation of document level and sentence level sentiment analysis also addressed in this research paper.

[11]. Pang, B., & Lee, L. (2019). Opinion mining and sentiment analysis. Foundations and Trends® inInformation Retrieval, 2(1-2), 1-135.

[12]. Cambria, E., & Hussain, A. (2019). Sentic computing: Techniques, tools, and applications. Springer.

[13]. Mohammad, S. M., & Turney, P. D. (2019). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

[14]. Liu, B. (2019). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

[15]. Thelwall, M., Buckley, K., & Paltoglou, G. (2018). Sentiment in Twitter events. Journal of the Association for Information Science and Technology, 63(1), 119-132.

[16]. Kouloumpis, E., Wilson, T., & Moore, J. D. (2018). Twitter sentiment analysis: The good the bad and theOMG!. ICWSM, 11(538-541), 164-167.

[17]. Pak, A., & Paroubek, P. (2018). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10).

[18]. Zhang, L., & Wu, D. (2017). A comprehensive survey on sentiment analysis of social media. Advances in Mechanical Engineering, 11(8), 1687814019867565.

[19]. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2017). Sentiment analysis of Twitter data. Proceedings of the Workshop on Languages in Social Media, 30-38.

[20]. Medhat, W., Hassan, A., & Korashy, H. (2017). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.