NORTH SOUTH UNIVERSITY

Sentiment Analysis of Bangladeshi Social media followers using Machine learning approach

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL AND COMPUTER ENGINEERING
OF NORTH SOUTH UNIVERSITY
IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Date

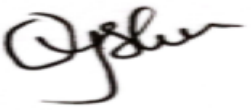
29th December 2022, Tuesday

Declaration

It is hereby acknowledged that:

- No illegitimate procedure has been practiced during the preparation of this document.
- This document does not contain any previously published material without proper citation.
- This document represents our own accomplishment while being Undergraduate Students in the **North South University**

We declare that this CSE498R report entitled *Sentiment Analysis of Bangladeshi Social media followers using Machine learning approach* has not been accepted for any degree and is not concurrently submitted in candidature of any other degree. We would like to request you to accept this report as a partial fulfillment of Bachelor of Science degree under Electrical and Computer Engineering Department of North South University. Sincerely,



Student 1: Shukdev Datta
1911838042

Student 2: Oyshee Rahaman
1611872042

Approval

This is to certify that the CSE498R report entitled Sentiment Analysis of Bangladeshi Social media Influencer of Facebook using Machine learning approach, submitted by Shukdev Datta (Student ID: 1911838042) and Oyshee Rahaman (Student ID: 1611872042) are undergraduate students of the Department of Electrical Computer Engineering, North South University. This report partially fulfils the requirements for the degree of Bachelor of Science in Computer Science and Engineering on October 25, 2022, and has been accepted as satisfactory.

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Dr. Md Shariful Islam
Assistant Professor
Department of Mathematics and Physics
North South University, Dhaka
Bangladesh.

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Dr. Rajesh Palit
Professor & Chair
Department of Electrical Computer Engineering
North South University, Dhaka
Bangladesh.

Abstract

Sentiment analysis is a technique that allows us to analysis the emotion and opinions from text data. The text data can be in the form of tweets, product reviews, status, posts etc. This technique allows us to find out the sentiment behind a text data and do further analysis. Usually sentiment analysis on tweeter dataset is very common practice but there had been no work on Sentiment analysis of Bangladeshi Influencers. It helps to correctly analyze the opinion of crowds when they view any post or status and they share their opinions in the form of text comments. This research work aims to proposed a method through which sentiment analysis can be carried out on the review dataset collected from Facebook page of Bangladeshi social media influencers. We have proposed a sentiment analysis technique that will do all the preprocessing required for the raw text data to be cleaned and then we have used TF-IDF vectorizer as the feature extraction technique to extract features. Label encoding was also performed along with text lemmatization in order to process the dataset to a clean noise free dataset. After all that preprocessing, we have applied various ML models and hyper-parameter tuning in order to classify reviews as "positive" or "negative". The results showed that the Random forest, Randomizer SearchCV, Grid SearchCV and Random SearchCV Model Specific scored 97 percent test accuracy, which is a very good accuracy on both the dataset that we have used.

Contents

List of Figures

List of Tables

Glossary

Logistic Regression :
Logistic Regression is one of the most important machine learning classifiers. Although the name suggests regression, but it is actually used to solve classification problems. Our project is a classification-based problem. Therefore, we have used logistic regression.

Logistic Regression L2 Regularization :
When we train a model, it is very easy for a model to overfit or underfit. In order to avoid such occurrences, we will use L2 regularization. It causes the model to be ideal and ensures model does not overfit.

Perceptron Model :
Perceptron model is another model that is used for binary classification tasks. It is also used in artificial neural network.

Multinomial Naive Bayes :
Naive Bayes (NB) is an algorithm that is commonly used in Supervised Learning and mainly used for solving classification problems. It is frequently used in text classification problems which include a large dataset with high dimensionality.

Random Forest Classifier :
Random Forest is an ensemble classifier and is made up of several decision trees applied to different subsets of the input dataset. It uses the majority voting for classification and the average of all the decision tree outputs.

SGD Classifier :
In SGD, a few of the samples will be selected randomly instead of using the whole dataset for each iteration.

Randomized SearchCV :
Randomized SearchCV is a type of hyper-parameter tuning technique, which is used when we have to try with many parameters and the training time of the model is very long.

Grid SearchCV :
It is an effective technique that can be used to adjust the parameters in such a way that

the cost function is minimum and improve the performance of the model.

Randomizer SearchCV Model Specific :

In this technique, we have used SVC as the model to be implemented in the randomized search hyper-parameter technique.

1 Introduction

1.1 Sentiment analysis

Sentiment analysis had always been a hot topic for researchers nowadays. Sentiment is a crucial part of human life and you cannot imagine a person without expressing sentiment. Sentiment analysis is a hot topic in Natural Language Processing. A technique is used to find out whether a statement is positive or negative. Using sentiment analysis expert, companies can easily understand the sentiment of their customers on their products and can improve the quality of those products. The novelty of our work is that there had been so many works on sentiment analysis but there had been no work on Sentiment analysis of Bangladeshi social-media influencers. The goal of our research work is to identify how the impression of the fans of these top social media influencers. So, in detail, we have collected social media reviews from the Facebook page of these social media influencers and stored them in our custom dataset. The models we have used are 1)Lorgistic Regression, 2)Logistic Regression L2 regularization, 3)Perceptron, 4)Multinomial Naïve Bayes, 5)Random Forest Classifier, 6)Stochastic Gradient Descent and hyper-parameter technique used are 1)Randomizer Search, 2)Grid Search and 3)Randomizer Search Model Specific(SVM). The results showed that the Random forest, Randomizer SearchCV, Grid SearchCV and Random SearchCV Model Specific scored 97 percent test accuracy, which is a very good accuracy on both the dataset that we have used.

1.2 Bangladeshi Social-media Influencers

The generation that we see nowadays is very different compared to the generation we saw in 90s or 80s. With the advancement of technology, came great internet connections, which created more social media platforms. The Bangladeshi kids are more stuck on their phone rather than playing outdoor games with their friends. The life has become so hectic that they have now thought of enjoying their break time watching these social media influencers perform singing, dancing, skating etc stuffs. This new generation is so obsessed with their favorite influencers that they tend to imitate what they do on live social media platform. These influencers act as the role model of these kids. For example, Shakib al Hasan, one of the most prominent cricketer of BD cricket has so many followers and people follow his way of playing and give positive impression when he plays well.

However, when his form is not there, the fan base starts to criticize his performance through negative remarks.

1.3 Literature review

Neethu et al.[5] have proposed a machine learning technique for their sentiment analysis research work that required preprocessing of their tweeter dataset. They have used Naive Bayes classifier, SVM classifier, Maximum Entropy Classifier and Ensemble classifier for their work. The accuracy they have achieved is the highest in SVM, Maximum Entropy, Ensemble, which was 93%. However, the accuracy scored by the Naive Bayes Classifier was 91%, which was least among all the four models used.

Rajkumar et al.[2] have proposed a sentiment analysis technique where they have collected data from Amazon in the form of json and the json files contained the reviews regarding different electrical appliances. As usual, they had to do some preprocessing before loading the dataset for training the ML models. They have used score generation, which is a technique that compares the words from the dataset with opinion lexicons and calculates the sentiment scores. Then they have used SVM and Naive Bayes in the form of models to train the dataset. SVM and Naive Bayes both scored the highest accuracy of 98.17% (Naive Bayes) and 93.54% (SVM) for camera dataset. The Naive Bayes model scored the least accuracy in TV dataset, which was 90.16%. SVM scored the least accuracy in Video surveillance dataset, which was 79.43%.

Bac Le et al.[3] have used twitter dataset for their sentiment analysis research work. As usual, they have also preprocessed their data for example, removing the stopwords, and removed unnecessary punctuations, urls etc. In case of feature extraction, the have used Bi-gram, Unigram and Object-Oriented feature extraction technique to collect features. They have used two models, which are SVM and Naive Bayes. SVM scored 80% accuracy and Naive Bayes scored 79.5% in their research proposed work.

Gautam et al.[1] have used twitter dataset in their research work. The dataset required preprocessing which includes removal of the repeated words and punctuation and improves the efficiency the data. They have used four methods, which are Naive Bayes, Maximum Entropy, Support Vector machine and Semantic Analysis (WordNet). Among all the methods used, Semantic Analysis (WordNet) have scored the accuracy of 89.9%, which was highest, and least was 83.8% for Maximum Entropy.

Palak et al.[4] have used IMDB movie review dataset which contained 1000 positive and 1000 negative reviews. During the preprocessing stage, the data is cleaned in order to remove the noise. Then tokenization is performed to extract features. They have used models, which are Naive Bayes, K Nearest Neighbor and Random Forest. Naive Bayes have scored the highest accuracy of 81.4% and K-nearest neighbor have scored the lowest accuracy of 55.30%.

2 Methodology

2.1 Data set

During the course of our research work we have used two dataset collected by us. The datasets include Athletes reviews and Motivational Speakers reviews. Each of these type of social media influencers will have own dataset of their own making it 2 specific dataset each category. All the data we collected for these 2 categories are from Bangladeshi celebrities. The dataset contains two columns, which are "review" and "sentiment". For each of the categories we have chosen 4 to 5 celebrities related to the two categories and collected fan reviews from their Facebook page and our sentiment column is to label those reviews as "positive" or "negative". Sample of two of the dataset is given below:

review	**sentiment**
Keep up the good work	positive
I wish someday Bangladesh Win the Fifa World Cup	positive
Big fan of Jamal Bhuyan	positive
Someday Bangladesh will play in Fifa World Cup	positive

Table 2.1: Athletes Dataset

review	**sentiment**
Thanks for this information	positive
Great advice!	positive
Thanks for this information	positive
Thank you ayman vaiya you are great	positive

Table 2.2: Motivational Speakers Dataset

2.2 Data Cleaning and Preprocessing

During our course of research work, we had to handle data pre-processing and data cleaning. It was very necessary to clean the data collected because it contains unnecessary characters that can affect the accuracy of the models during training. Therefore, we have to do data cleaning so that we get good accuracy of the models. During cleaning stage, we will look for hash, '@', 'http', stopwords, HTML tags, special characters and single characterswhich needs to be removed. We have created function to remove these things and once the reviews are cleaned, we place the cleaned review in a new column in the data-frame.

2.3 Feature Selection

After cleaning the whole dataset, the samples were preprocessed properly. At first, we started the text lemmatization. Lemmatization is a process that is used to find words with similar meaning like democracy, democratic, and democratization and transform them into democracy everywhere in the text and we will only need one dimension to represent these multiple words. Next, we have to use label encoding to pass the labels into the models since the models cannot understand textual words like "positive" and "negative". Label encoding will transform "positive" into 0 and "negative" into 1. We then used train-test split where training set is 75 percent of the whole dataset and test set is 25 percent of the whole dataset.

We have used TF-IDF Vectorizer for feature transformation. This technique will be used to transform text into meaningful representation of numbers, which will be used in machine learning algorithms for training. TF means how often a specific team appears in the whole document and IDF means how common or rare a term is across the entire corpus of documents.

2.4 Machine Leaning Classifiers and Hyper-parameter tuning

In order to classify the sentiment of the reviews collected from the social influencer dataset (Athletes and Motivational Speakers), we have used some very efficient machine learning classifiers. In order to explore the performance of our work, these models will be a great tool to identify the performance measures like accuracy, precision, recoil, f1 score etc. The ML models used are given below:

- Lorgistic Regression
- Lorgistic Regression (L2 regularization)
- Perceptron
- Multinomial Naive Bayes
- Random Forest Classifier
- Stochastic Gradient Classifier (SGD)

The hyper-parameter technique used are given below:

- Randomized Search
- Grid Search
- Randomized Search Model Specific (SVM)

Our dataset contains two features. One of them is stored as X and another one of them is stored as Y. Here Y is the target and X is the feature. X will contain all the "review" data and Y will contain all the "sentiment" data. Since we have used various machine-learning models, so our primary choice of performance measure is "accuracy". Now let's talk more about our ML models and hyper parameter technique that we have used.

2.4.1 Logistic Regression

Logistic Regression is one of the most important machine learning classifiers. Although the name suggests regression, but it is actually used to solve classification problems. Our project is a classification-based problem. Therefore, we have used logistic regression. Logistic regression does not give direct 0 or 1 as output but rather it gives out in the form of a probability in between 0 and 1. The basic equation of logistic regression is given below:

$$y = by = b_0 + b_1x_1 + \cdots + b_nx_n$$

Logistic Regression Equation

2.4.2 Logistic Regression L2 Regularization

When we train a model, it is very easy for a model to overfit or underfit. In order to avoid such occurrences, we will use L2 regularization. It causes the model to be ideal and ensures model does not overfit. If we let our model to overfit, then the model will be biased. This is usually useful when dealing with polynomial data. It adds a penalty term with the cost function, which causes the decision boundary less complicated, and model can avoid overfitting.

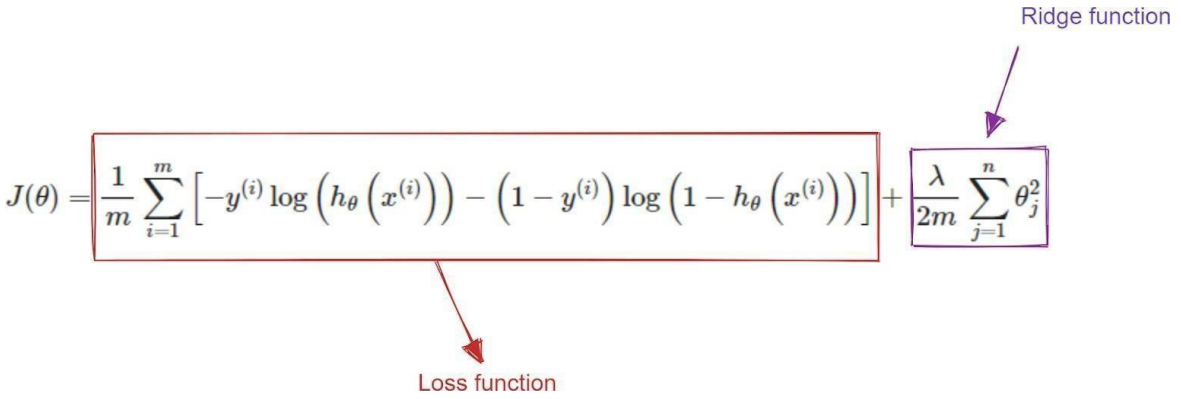


Figure 2.1: Logistic Regression L2 regularization Equation

2.4.3 Perceptron Model

Perceptron model is another model that is used for binary classification tasks. It is also used in artificial neural network. It is a single-layer neural network consisting of four main parameters, which are: input values, weights and Bias, net sum, and an activation function. The working principle of perceptron is to multiply the input values with the weight and then concatenate these results to produced weighted sum and pass it to activation function in order to obtain the final output.

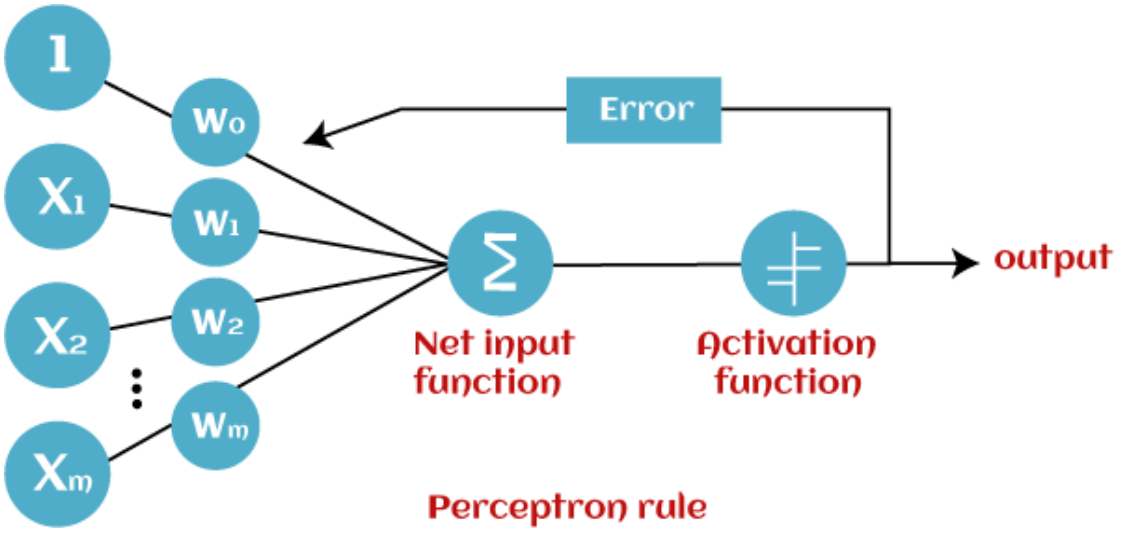


Figure 2.2: Logistic Regression L2 regularization Equation

2.4.4 Multinomial Naive Bayes

Naive Bayes (NB) is an algorithm that is commonly used in Supervised Learning and mainly used for solving classification Problems. It is frequently used in text classification problems, which include a large dataset with high dimensionality. NB is aprobabilistic classifier, which means that it predicts the target based on the probability of occurrence. In this study, we have used the Multinomial Naive Bayes (MNB).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Multinomial Naïve Bayes

2.4.5 Random Forest Classifier

RF is an ensemble classifier and is made up of several decision trees applied to different subsets of the input dataset. It uses the majority voting for classification and the average of all the decision tree outputs.

2.4.6 SGD Classifier

In SGD, a few of the samples will be selected randomly instead of using the whole dataset for each iteration. SGD is extremely useful when you are dealing with a huge dataset. SGD is an efficient optimization algorithm that is used to find the values of parameters for which the cost function is minimum.

2.4.7 Randomized Search

Randomized SearchCV is a type of hyper-parameter tuning technique, which is used when we have to try with many parameters and the training time of the model is very long. Define a search space as a bounded domain of hyper-parameter values and randomly sample points in that domain.

2.4.8 Grid SearchCV

Define a search space as a grid of hyperparameter values and evaluate every position in the grid. It is an effective technique that can be used to adjust the parameters in such a

way that the cost function is minimum and improve the performance of the model.

2.4.9 Randomized Search Model Specific

In this technique, we have used SVC as the model to be implemented as the randomized search hyper-parameter technique.

2.5 Flow Chart

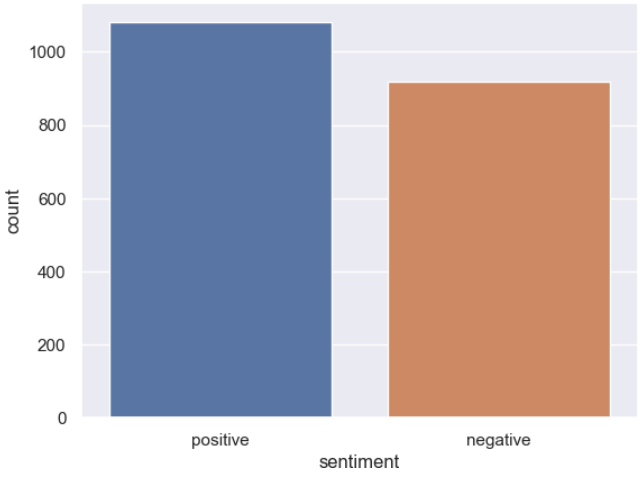
(a) Principle Component Analysis (PCA) for PD and control samples (2D plotting)

Figure 2.5: PCA Plotting

3 Results and Analysis

3.1 For Athletes' Dataset:

Athletes' dataset contained the positive and negative reviews of some of the top popular athletes of Bangladesh. Out of 1999 reviews collected, we have found out that there are 1080 positive reviews and rest are all negative reviews. Therefore, if we want to say the distribution of positivity and negativity that people feel about the athletes are 54.02% of fans are positive and 48.98% of fans are negative on their performance in sports.



(a) Histogram representing the visualization of positive and negative sentiment reviews

Figure 3.1: Count VS Sentiment Histogram Plot for Athletes Dataset

In Table 3.1, we demonstrated the performance matrices for the highest accuracy values for athletes' dataset.

In the table 3.1 below, we have seen that the Random Forest, Randomizer SearchCV, Grid SearchCV and Random SearchCV Model Specific all have same accuracy of 97%, which is the highest among all the other models. The second highest accuracy is for Perceptron and Stochastic Gradient Descent, which is 96%, and least accuracy is for Logistic Regression L2 Regularization and Multinomial Naive Bayes is 93%.

Model Name	Accuracy	Precision	Recall	F1-score
Logistic regression	0.96	0.96	0.96	0.96
Logistic Regression L2 regularization	0.93	0.93	0.93	0.93
Perceptron	0.96	0.96	0.96	0.96
Multinomial Naive Bayes	0.93	0.93	0.93	0.93
Random Forest	0.97	0.97	0.97	0.97
Stochastic Gradient Descent	0.96	0.96	0.96	0.96
Randomized SearchCV	0.97	0.97	0.97	0.97
Grid SearchCV	0.97	0.97	0.97	0.97
Random SearchCV Model Specific	0.97	0.97	0.97	0.97

Table 3.1: Model Performance Matrices on Athletes Dataset

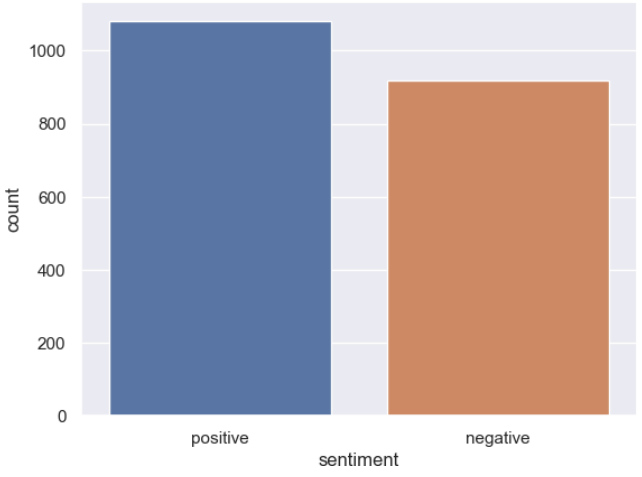
3.2 For Motivational Speaker Dataset:

Motivational Speaker dataset contained the positive and negative reviews of some of the top popular motivational speaker of Bangladesh. Out of 1999 reviews collected, we have found out that there are 1080 positive reviews and rest are all negative reviews. Therefore, if we want to say the distribution of positivity and negativity that people feel about the motivational speaker are 54.02% of fans are positive and 48.98% of fans are negative on their podcasts.

In Table 3.2, we demonstrated the performance matrices for the highest accuracy values for Motivational Speaker dataset.

Model Name	Accuracy	Precision	Recall	F1-score
Logistic regression	0.96	0.96	0.96	0.96
Logistic Regression L2 regularization	0.93	0.93	0.93	0.93
Perceptron	0.96	0.96	0.96	0.96
Multinomial Naive Bayes	0.93	0.93	0.93	0.93
Random Forest	0.97	0.97	0.97	0.97
Stochastic Gradient Descent	0.96	0.96	0.96	0.96
Randomized SearchCV	0.97	0.97	0.97	0.97
Grid SearchCV	0.97	0.97	0.97	0.97
Random SearchCV Model Specific	0.97	0.97	0.97	0.97

Table 3.2: Model Performance Matrices on Motivational Speaker Dataset



(a) Histogram representing the visualization of positive and negative sentiment reviews

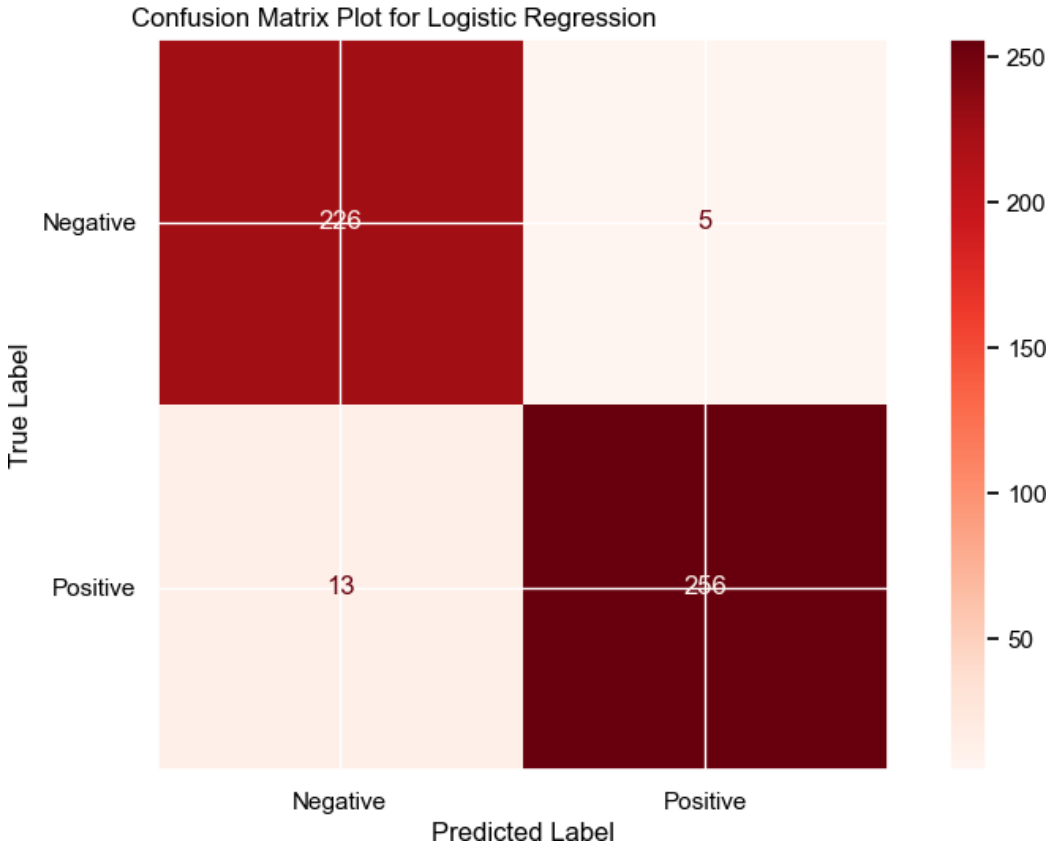
Figure 3.2: Count VS Sentiment Histogram Plot for Motivational Speaker Dataset

In the table 3.2 above, we have seen that the Random Forest, Randomizer SearchCV, Grid SearchCV and Random SearchCV Model Specific all have same accuracy of 97%, which is the highest among all the other models. The second highest accuracy is for Perceptron and Stochastic Gradient Descent, which is 96%, and least accuracy is

for Logistic Regression L2 Regularization and Multinomial Naive Bayes is 93%.

3.3 Confusion Matrix and ROC Curve:

3.3.1 Logistic Regression:



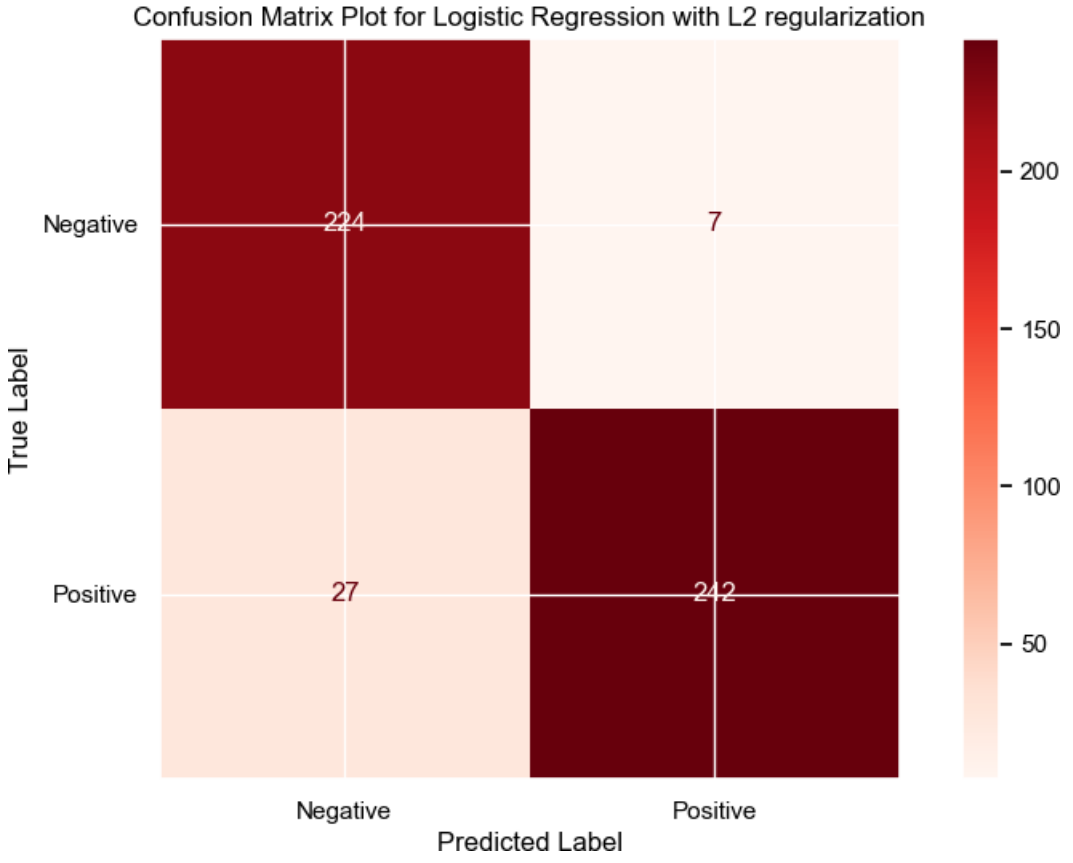
(a) Confusion Matrix for Logistic Regression

Figure 3.3: Confusion Matrix showing performance measure of LR model

The confusion matrix shows that there are 256 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Postive rate is equal to TP divided by actual "positive". Hence true positive rate is (256)/(256+13) which will be 0.952. There are 226 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (226)/(226+5) which will be 0.978. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (13+5)/(231+269) which will be 0.036.

3.3.2 Logistic Regression L2 Regularization:

The confusion matrix shows that there are 242 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Positive rate is equal to TP divided by actual "positive". Hence true positive rate is (242)/(242+27) which will be 0.8996. There are 224 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (224)/(224+7) which will be 0.970. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (27+7)/(231+269) which will be 0.068.



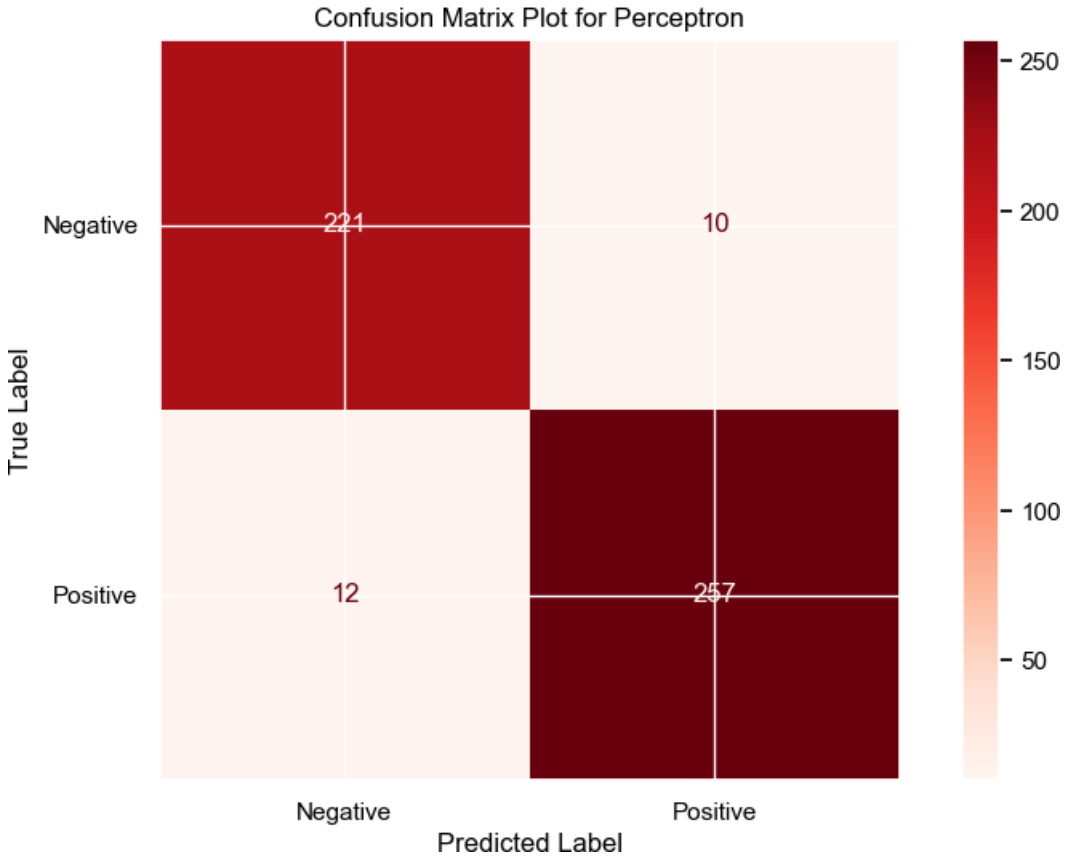
(a) Confusion Matrix for Logistic Regression L2 regularization

Figure 3.4: Confusion Matrix showing performance measure of LR L2 model

3.3.3 Perceptron:

The confusion matrix shows that there are 257 TP cases, w h i c h means that the model predicts "positive" and the label actually was "positive". True Positive rate is equalto TP

divided by actual "positive". Hence true positive rate is (257)/(257+12) which will be 0.955. There are 221 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (221)/(221+10) which will be 0.957. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (12+10)/(231+269) which will be 0.044.



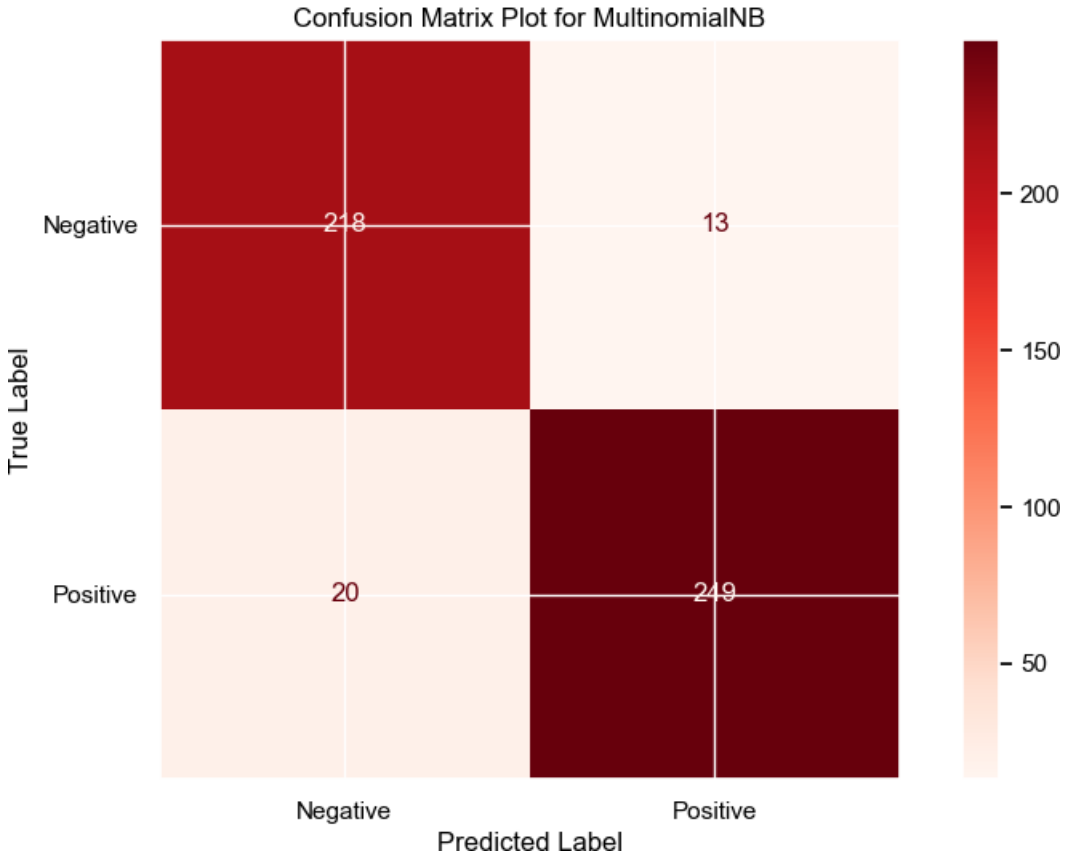
(a) Confusion Matrix for Perceptron

Figure 3.5: Confusion Matrix showing performance measure of perceptron model

3.3.4 Multinomial Naive Bayes:

The confusion matrix shows that there are 249 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Positive rate is equal to TP divided by actual "positive". Hence true positive rate is (249)/(249+20) which will be 0.926. There are 218 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (218)/(218+13) which will be 0.944. Total misclassification rate

will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (20+13)/(231+269) which will be 0.066.

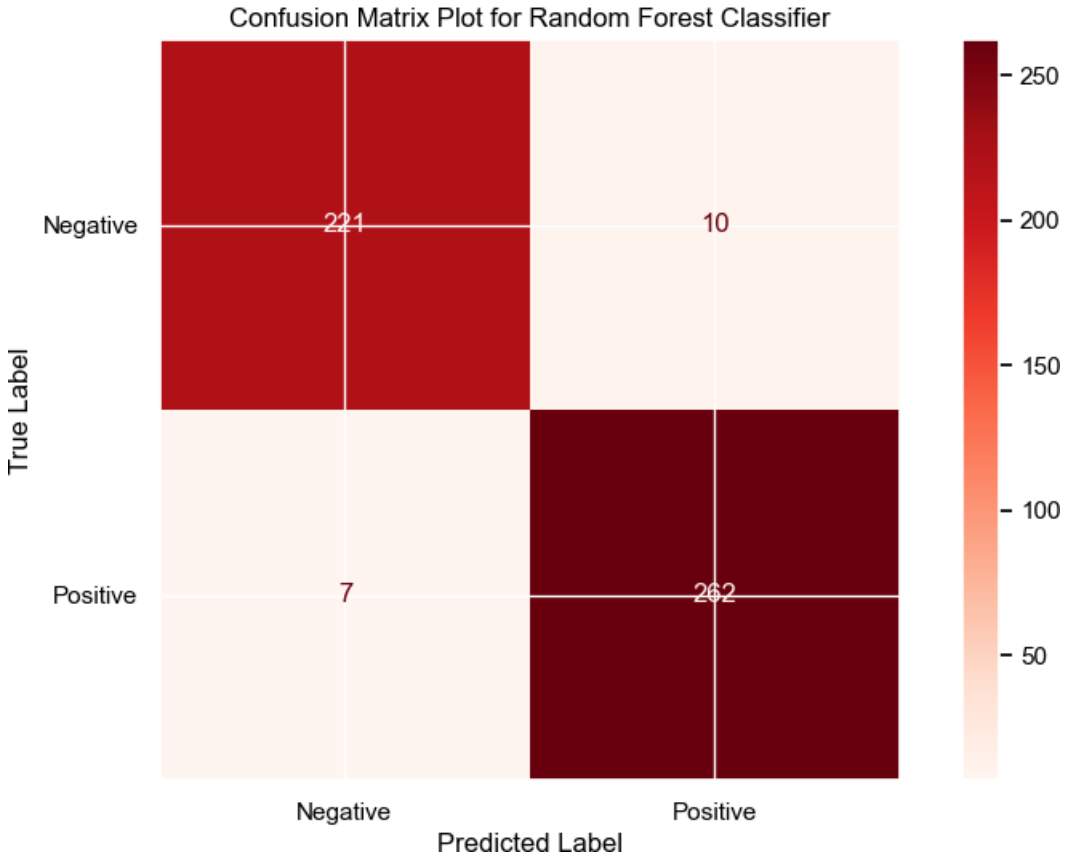


(a) Confusion Matrix for Multinomial Naive Bayes

Figure 3.6: Confusion Matrix showing performance measure of Multinomial Naive Bayes model

3.3.5 Random Forest:

The confusion matrix shows that there are 262 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Positive rate is equal to TP divided by actual "positive". Hence true positive rate is (262)/(262+7) which will be 0.974. There are 221 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (221)/(221+10) which will be 0.957. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (10+7)/(231+269) which will be 0.034.

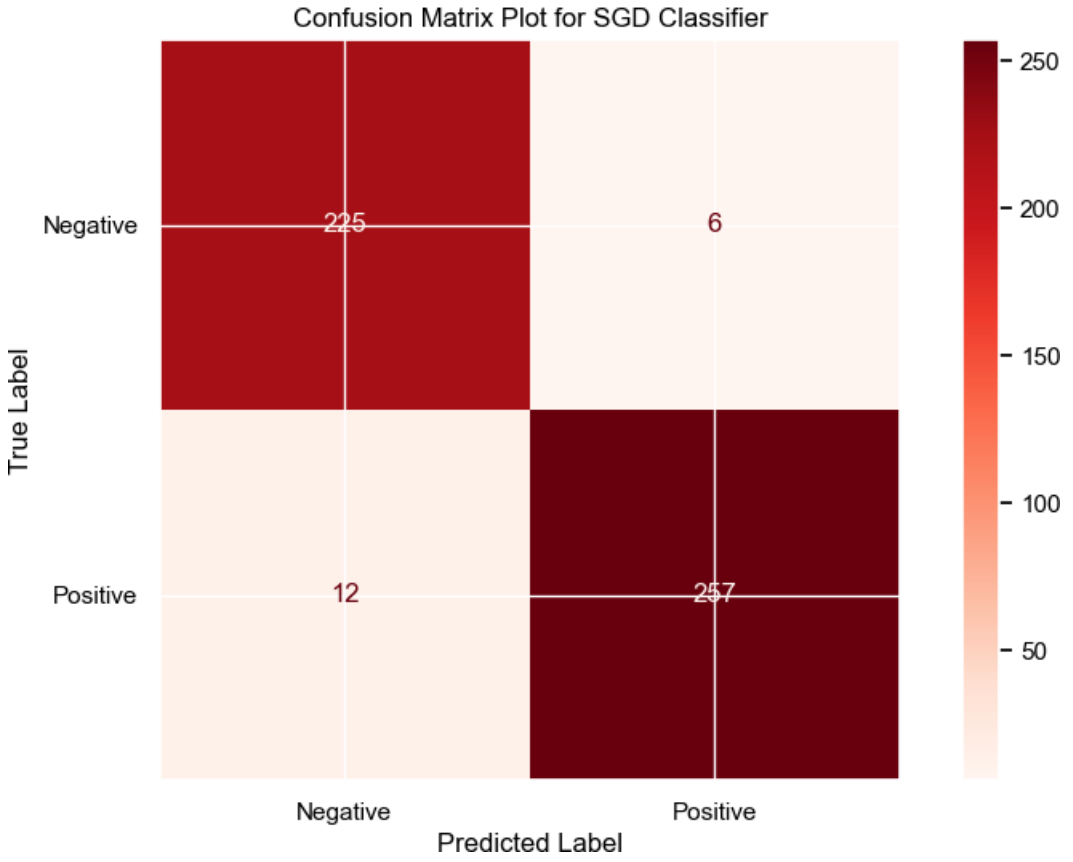


(a) Confusion Matrix for Random Forest

Figure 3.7: Confusion Matrix showing performance measure of Random Forest model

3.3.6 Stochastic Gradient Descent Classifier:

The confusion matrix shows that there are 257 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Positive rate is equal to TP divided by actual "positive". Hence true positive rate is (257)/(257+12) which will be 0.955. There are 225 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (221)/(221+10) which will be 0.974. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (12+6)/(231+269) which will be 0.036.

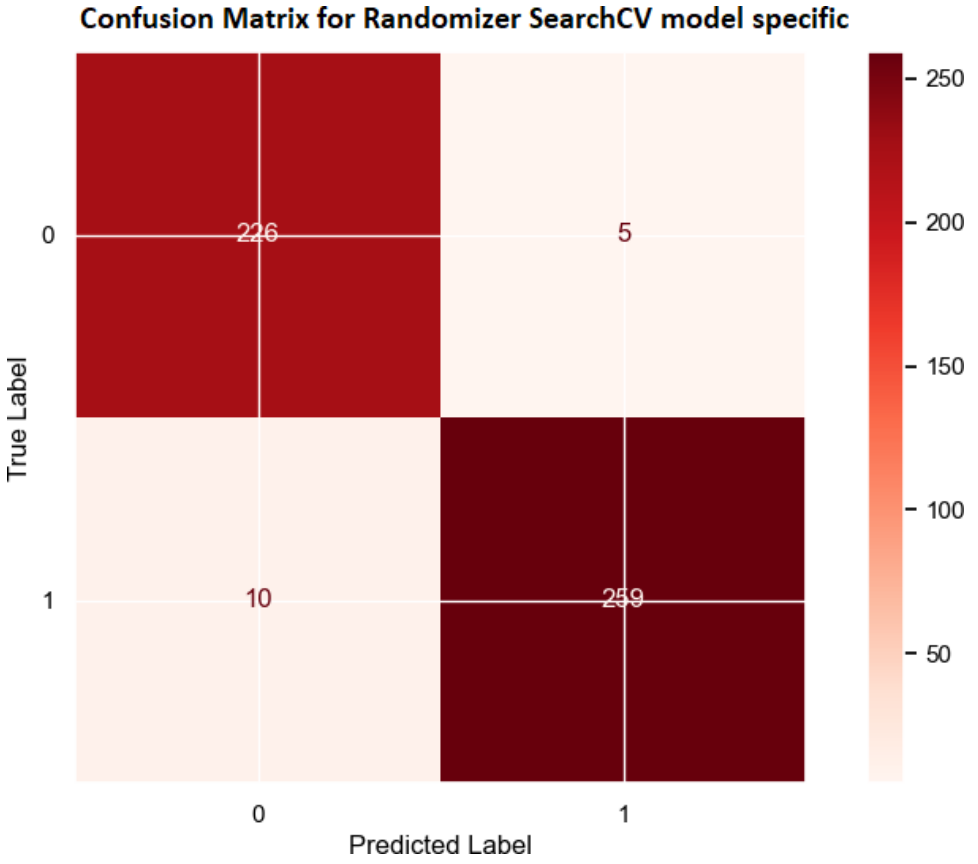


(a) Confusion Matrix for Stochastic Gradient Descent Classifier

Figure 3.8: Confusion Matrix showing performance measure of Stochastic Gradient Descent Classifier model

3.3.7 Randomized SearchCV Model Specific:

The confusion matrix shows that there are 259 TP cases, which means that the model predicts "positive" and the label actually was "positive". True Positive rate is equal to TP divided by actual "positive". Hence true positive rate is (259)/(259+10) which will be 0.963. There are 226 TN cases where the model predicted "negative" and the actual label was "negative". True negative rate is equal to TN divided by actual "negative". Hence the True negative rate is (226)/(226+5) which will be 0.978. Total misclassification rate will be equal to (FN+FP)/Total number of cases. Hence misclassification rate will be (10+5)/(231+269) which will be 0.030.

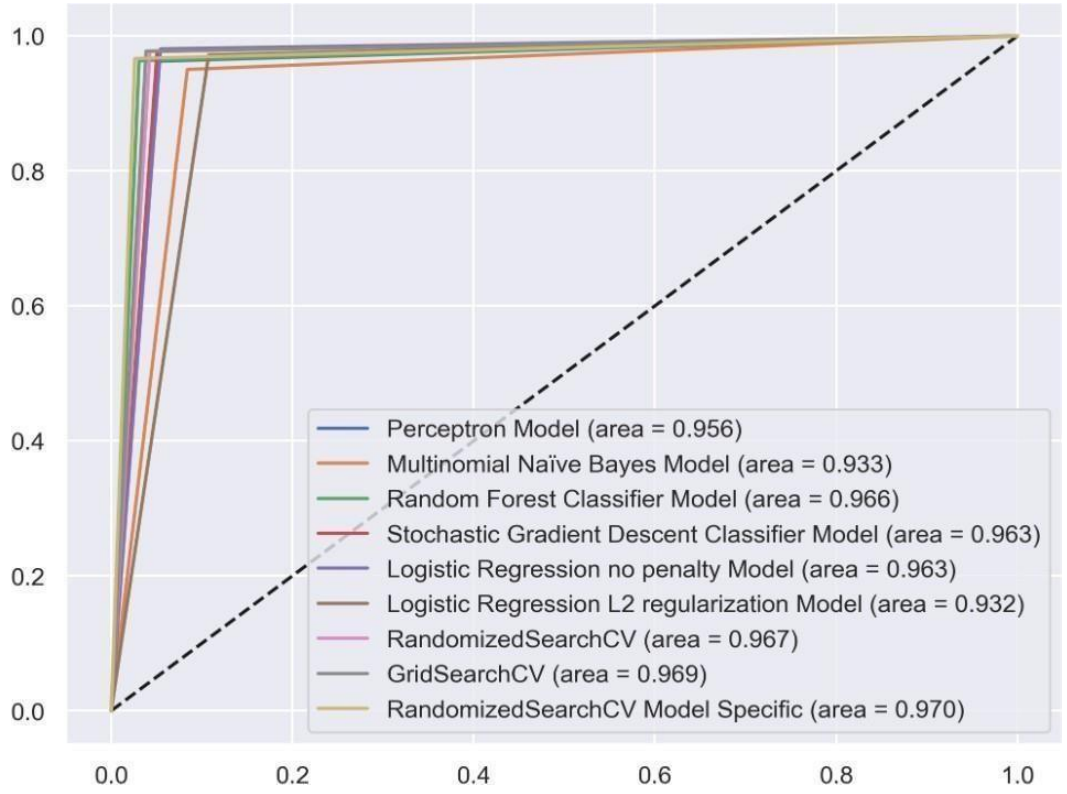


(a) Confusion Matrix for Randomized SearchCV Model Specific

Figure 3.9: Confusion Matrix showing performance measure of Randomized SearchCV Model Specific model

3.3.8 ROC Curve:

The ROC curve shows that among all the models applied in our research work, Randomized SearchCV model specific has the highest accuracy of 97% and Logistic Regression L2 Regularization model has the least accuracy of 93.2%.



(a) ROC Curve

Figure 3.10: ROC Curve for all models

4 Conclusion

4.1 Discussion and conclusion

In the present study, we have used our two custom dataset that have been collected by ourselves. It adds novelty to our work since the dataset is ours and no work has been done on such topic yet. We are the first research group to introduce this kind of work. We believe that it will create opportunities for future data science research groups to work more on our works and develop new and fresh ideas taking help from our research paper. The preprocessing phase was very essential since the raw data contained great deal of noises that could make the model learn very less and the performance measures like accuracy would have been hampered. We have used number of models to actually observe how these models perform in the custom dataset of ours. The impact of the social media influencers of Bangladesh have on the new generation have also been taken account of. It is observed that based on recent events the fans of these social media influencers express positive or negative remarks on the verified Facebook pages of these influencers. Moreover, our goal was to identify how much positive or negative these fans are based on their influencer's work.

4.2 Future Work

In this current study, we have used the ML classifier models and hyper-parameter tuning techniques to classify a statement as "positive" or "negative". However, we actually want to increase the content of our dataset from 1999 reviews to say 10000 reviews. In this case, we may actually need the use of deep learning since deep learning have become a common trend nowadays when dealing with large datasets. In addition, the performance of deep learning models is even better that ML Classifier models. So, in future work, we will actually increase the dataset size and introduce Deep Learning models to do training on our new dataset. We also plan to add more category of followers of influencers where initially in this research work; we have only used two, which are "Athletes" and "Motivational Speaker". In addition, we would like to add the data of the influencer in future to perform sentiment analysis on both the influencer and the followers.

References

[1] Gautam, G. a. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)* (pp. 437-442). IEEExplore. doi:10.1109/IC3.2014.6897213

[2] Jagdale, R. S. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive informatics and soft computing* (pp. 639--647). Springer.

[3] Le, B. a. (2015). Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering* (pp. 279--289). Springer.

[4] Palak Baid, A. G. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. *International Journal of Computer Applications, 179*(7), 45-49. doi:10.5120/ijca2017916005

[5] Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine-learning techniques. In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT) (pp. 1-5). IEEE.