PRML PROJECT

IIT JODHPUR

Twitter Sentiment Analysis

Made by

Mayank Singh Rajput(B19CSE054),Mohit Ahirwar(B19CSE055),Ram Khandelwal(B19CSE116)

I. Introduction

The problem statement expects us to perform twitter sentiment analysis by using various classifiers and calculating accuracies for each of them and comparing the results.The classifiers predicts the target value for a tweet statement after it has been trained on the training data.The target value depicts a particular sentiment of a tweet present in the statement.

II. Observing the Dataset

The dataset consists of 6 columns which include id,date,query,username,content and target.While observing the content column we could see that various contractions,emojis,urls,stopwords and other texts are present.This requires us to move to the next step of text preprocessing.

III.Text Preprocessing

Following are the text preprocessing steps mentioned.

A. Replacing contractions

Contractions are short forms of some words that are used in common communication english language.We have replaced them with their full forms so that it becomes easier to further analyse the data.

B. Defining and replacing emojis

Some emojis are positive which include smile,laugh,love and wink whereas some emojis are negative which include sad and cry emotions.Positive emojis are replaced with “positiveemoji” and negative emojis are replaced with “negativeemoji”.

C. Removing unwanted text

We have removed unwanted text which include usernames,URLs,digits,quotes,all single characters and punctuations.

D. Other Preprocessing

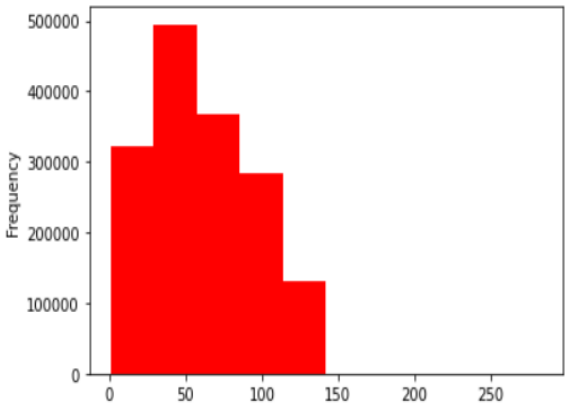
Other preprocessing include lower casing the string,replacing double spacing with single space,and converting more than 2 letter repetitions to 2 letter.

IV. Findings

A.Showing count of positive and negative emotions

The plot above shows that both positive and negative emotions are present in equal amounts in the dataset.

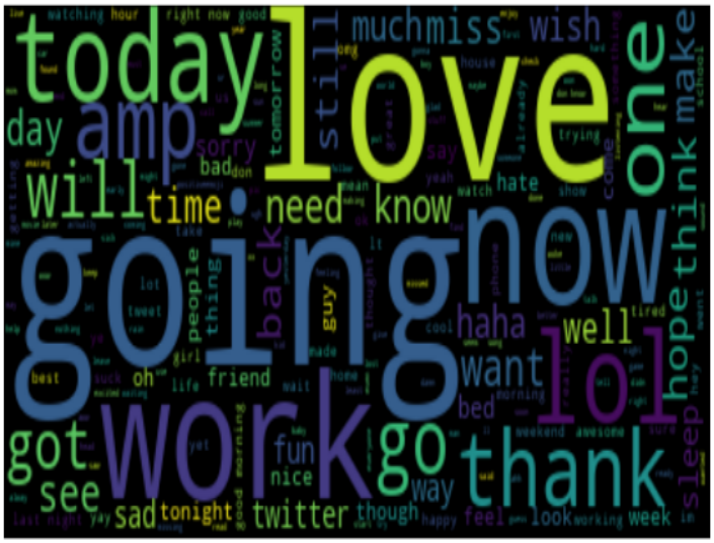
B.Showing Length of tweet vs frequency



The above plots depicts the relationship between the length of processed tweet with the frequency in the given dataset.It appears that most of the tweets are and around 50.

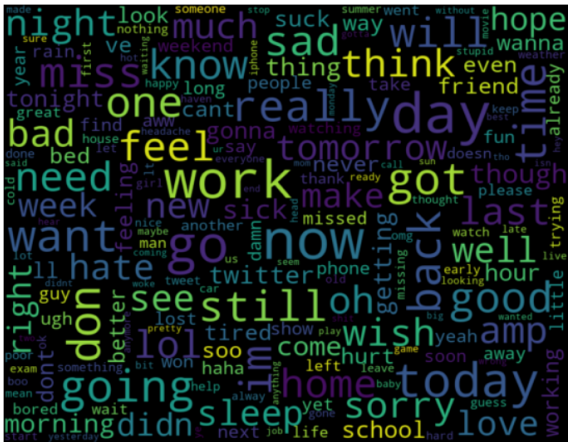
V.ANALYSIS USING WORD CLOUD

For full dataset-We could see that love,today,going,work,thank are more frequent.



For positive sentiments-We could see that thank,good,great,love,will are more frequent.

For negative sentiments -We could see that words such as feel,work,think,sad,hope,etc,are more frequent.



V.TEXT FEATURE EXTRACTION

For text feature extraction we have used CountVectorizer function imported from sklearn library.It implements "Bag Of words" strategy which uses tokenization,counting and normalization.Tokenization is the process by which an integer id is given to the strings separated by tokens which could be punctuations and white spaces.Counting counts the number of tokens present in the document.Normalizing is the process of transforming a text into canonical form.

VI. CLASSIFIERS USED

A.MULTINOMIAL NAIVE BAYES

Multinomial Naive bayes implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification.This classifier is imported from sklearn.naive_bayes.The model is fitted to this classifier and test data is predicted.

B.LINEAR SUPPORT VECTOR CLASSIFIER

Linear SVC uses svm algorithm which in turn uses linear kernel function.We have used one verses rest technique which splits the multi-class classification into one binary classification problem per class.We have taken C=1,square hinge loss.The model is fitted to this classifier and test data is predicted.

C.LOGISTIC REGRESSION

Logistic Regression is a regression technique that uses a logistic function to model a binary dependent variable.The model is fitted to this classifier and test data is predicted.

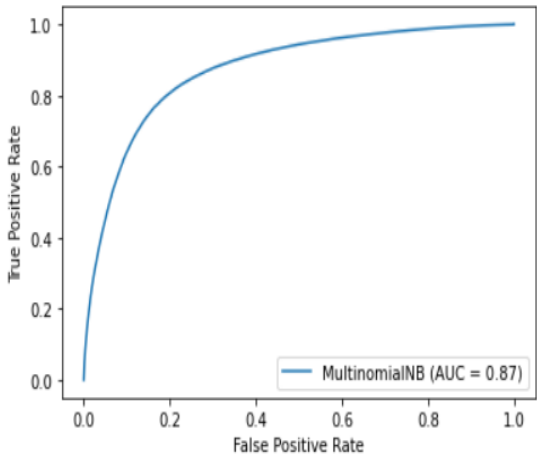
D.SVM

Support vector machine is a supervised learning technique used to classify data with the help of kernel functions which include linear,polynomial and gaussian functions.It is effective in use with high dimensional data.Various hyperparameters such as C value,gamma value,tolerance,coeff_, etc affect the accuracy produced.

VII. RESULTS AND COMPARISON

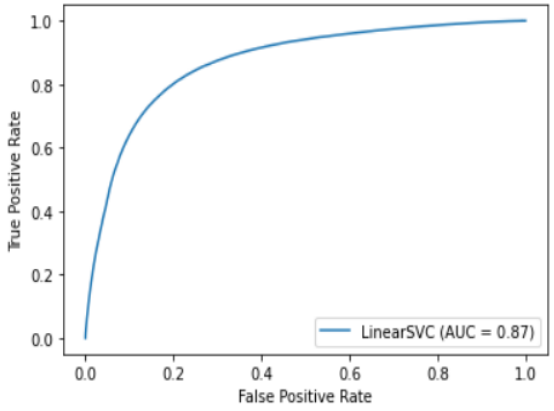
Note we have used
0 for-negative sentiment
4 for-positive sentiment

ROC plot and Confusion Matrix for Multinomial Naive Bayes



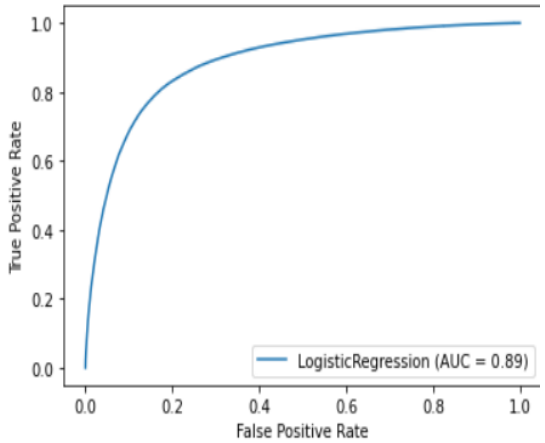
	precision	recall	f1-score	support
0	0.79	0.83	0.81	159808
4	0.82	0.78	0.80	160192
accuracy			0.80	320000
macro avg	0.80	0.80	0.80	320000
weighted avg	0.80	0.80	0.80	320000

ROC Plot for Linear SVC



	precision	recall	f1-score	support
0	0.81	0.79	0.80	159808
4	0.80	0.81	0.80	160192
accuracy			0.80	320000
macro avg	0.80	0.80	0.80	320000
weighted avg	0.80	0.80	0.80	320000

ROC Plot for Logistic Regression



	precision	recall	f1-score	support
0	0.83	0.79	0.81	159808
4	0.80	0.84	0.82	160192
accuracy			0.81	320000
macro avg	0.82	0.81	0.81	320000
weighted avg	0.82	0.81	0.81	320000

On comparing the above performance we see that
For 0 class

Comparison	Classifier
Precision	Logistic Regression>Linear SVC> Multinomial Naive Bayes
Recall	Multinomial Naive Bayes>Linear SVC=Logistic Regression
F1 score	Logistic Regression=Multinomial Naive Bayes>Linear SVC
AUC	Logistic Regression>Linear SVC= Multinomial Naive Bayes

For 4 class

Comparison	Classifier
Precision	Multinomial Naive Bayes>Linear SVC=Logistic Regression
Recall	Logistic Regression>Linear SVC> Multinomial Naive Bayes
F1 score	Logistic Regression>Linear SVC=Multinomial Naive Bayes
AUC	Logistic Regression>Linear SVC= Multinomial Naive Bayes

Below is the results of the accuracy results of all the three classifiers used above to predict the model.

Classifier	Accuracy
Multinomial Naive Bayes	80.20%
Linear SVC	80.07%
Logistic Regression	81.4675%

We see that Logistic Regression has performed better as compared to the other 2.Linear SVC and Multinomial had almost done equally better.

VIII.CONTRIBUTIONS

Here Coding Part :

- **Mayank** -Logistic Regression,SVM,Decision trees,Xgboost, Stopwords.
- **Mohit** - Multinomial Naive Bayes,Replacing emojis with their negative and positive sentiment,performance measurement and plots.
- **Ram** - Text preprocessing ,test-train split and count vectorizer,LinearSVC

CONCLUSIONS

On doing the sentiment Analysis, we found that there were equal number of positive and negative sentiment and no neutral sentiments were present.We used various classifiers for twitter Sentiment analysis and reached to the conclusion that Logistic Regression classifier was best of all which gave us the

highest accuracy.We followed machine learning pipeline step by step which helps us to systematize our approach towards solving problems for machine learning.

Acknowledgment

.We have successfully completed the twitter sentiment analysis by using various classifiers and comparing their accuracies.We also learnt how to use machine learning pipeline in a project from preprocessing,feature selection,learning a model to classification and finding accuracy scores.We got to learn a lot from this project.We express our gratitude to Dr Richa for giving us this opportunity to work on this project.

References

[1] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[4] https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
[5] https://scikit-learn.org/stable/modules/feature_extraction.html