Twitter Sentiment Analysis Project Report

Khushboo Singh, *B19EE046,* Chirag Bhawnani, *B19EE022* and Pragati Sinha, *B19CSE065*

Abstract

The abstract goes here.

Index Terms

IEEE, IEEEtran, journal, LATEX, paper, template.

I. Introduction

THIS file is the report for our project twitter sentiment analysis produced under LATEX using IEEEtran.cls version 1.8b and later. We wish you get complete gist of our approach, finding and results achieved while reading this report.

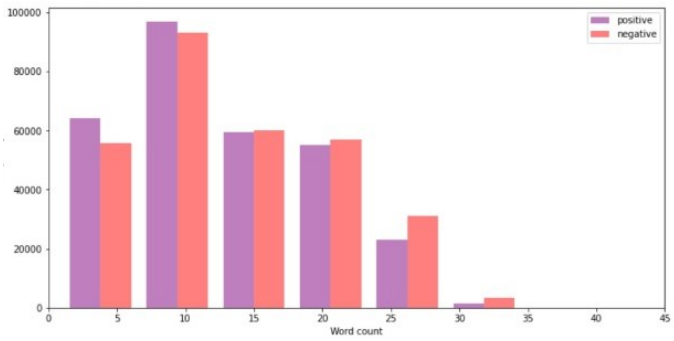
Team
May 16, 2021

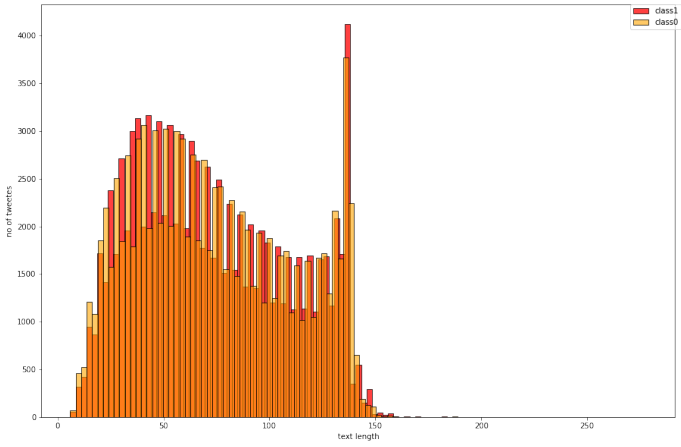
A. About the dataset

The dataset contains 1,600,000 rows and 6 columns.The first column shows the polarity of the tweets (binary class 0 and 4), 0 stands for negative and 4 stands for positive.The remaining 5 features are : id, date, query, user, text.

1) Approach: The analysis was initialized by encoding the polarity value 4 to 1.
The dataset was studied with respect to its distribution, missing values, number of positive and negative tweets as well as their correlation with the length and word count respectively.
The data was further processed by removing double spaces, urls, punctuations, stopwords, frequent words, rarewords and emojis etc. The cleansed data was visualized using wordcloud was made for tweets of each polarity.
Feature extraction matrix (TF-IDF) was created for information retrieval to represent how important a specific word or phrase was and the data was splitted into training and testing sets.
Different classification models like Logistic Regression, Random Forest, Naive Bayes, SVM and Perceptron were build to study their performance on the dataset.
The data was further analyzed in terms of most frequent hashtags in each respective polarity as well as the emotions association with them.

2) Findings: For lesser tweet length positive tweets were slightly higher than negative tweets.For lesser number of words positive tweets were more than negative tweets.Same analysis was made for 2 lakhs , 4 lakhs and 6 lakhhs dataset.

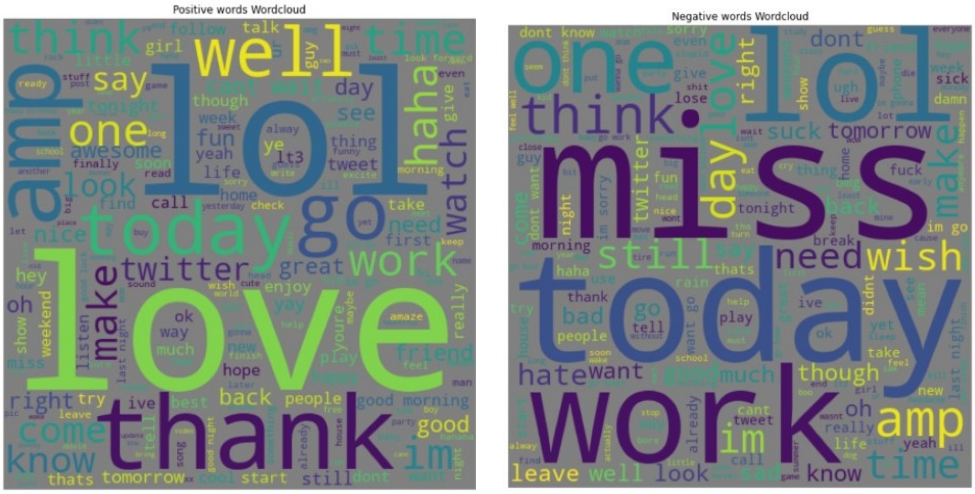




For positive word cloud words such as love,thank,well,lol were of bigger size that is common.For negative word cloud words

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).
J. Doe and J. Doe are with Anonymous University.
Manuscript received April 19, 2005; revised August 26, 2015.

like miss,work,lol,today,one were of bigger size that is higher frequency.The word lol appeared with higher frequency in both the word cloud.Hence proving we can't much rely on frequency of particular words for classification but a set of words would be required for it.Hence we have made TF-IDF matrix and taken each word as feature.



Upon training the model we got maximum accuracy on testing data set for logistic regression model and least from perceptron. LinearSvm gave very high accuracy in training data set but testing accuracy decreased compared to logistic regression means it is overfitting the data.

For random forest neither entropy nor gini method was able to give high accuracy.

Cross validation score for 5 validation and testing accuracy were almost the same.This means the models are not sensitive for the change in training , testing data.

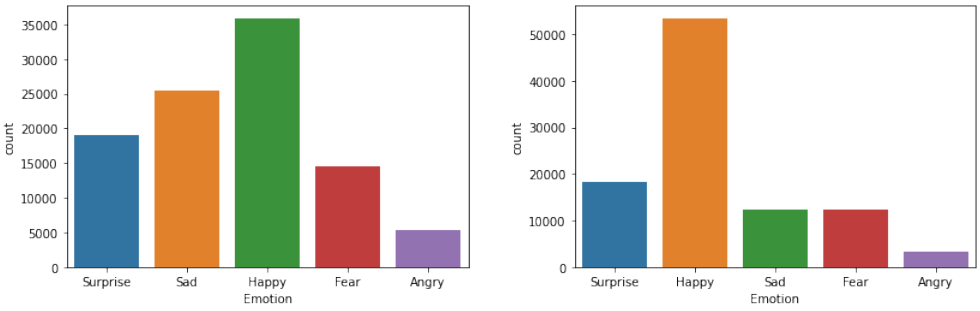
Naive bayes(multinomial naive bayes used) is predicting 40 percent true negative values which is highest among all means it is able to detect negative tweets best perfectly.Multinomial bayes is mostly used in Natural Language Processing and here also it does its work well.

Random forest has highest false positive it is classifying a lot of tweets of negative sentiment as positive.Random forest takes selected number of words at a time for model building thats why its happening. Random forest is not a good algorithm here.

Multimonial naive bayes has performed well for false positives as well it has very low false positives.

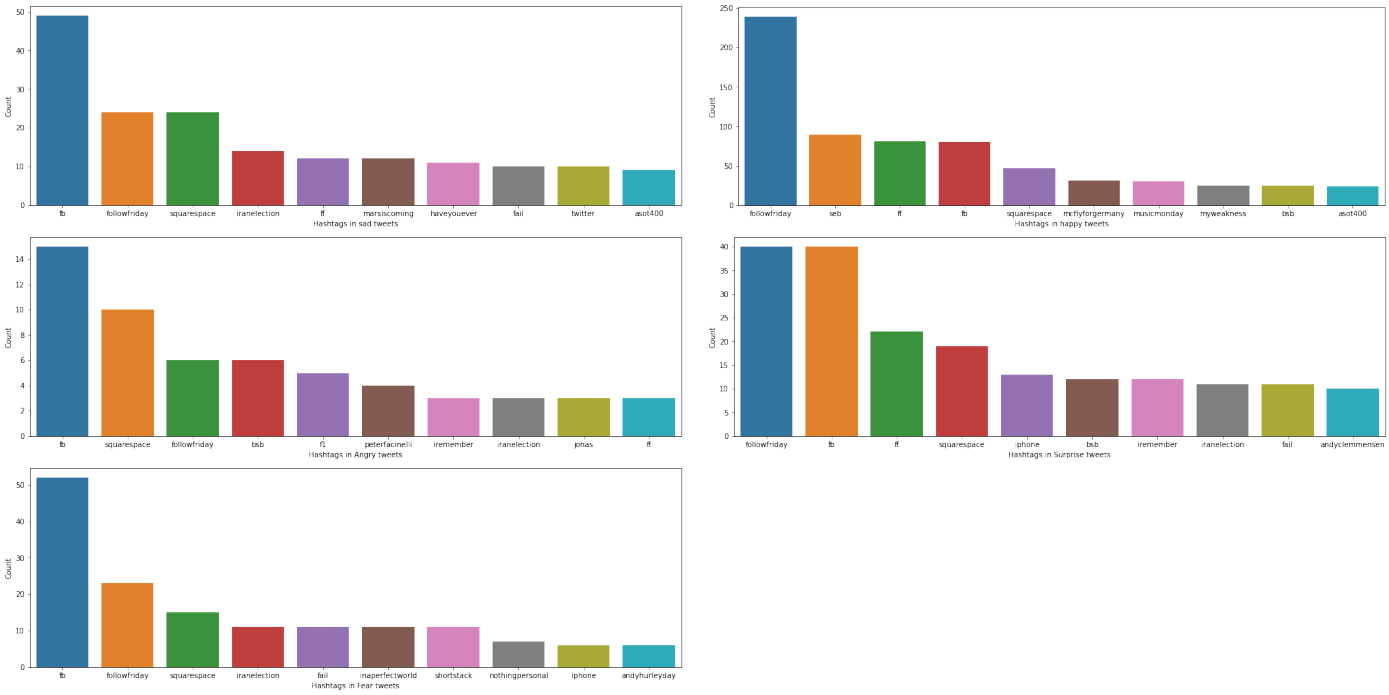
However multimonial naive bayes has higher value of false negative as well showing its biasness towards negative tag making it not a good classification algorithm here.

Perceptron also has high false negative,false positive and lower true prediction making it not a good algorithm to be used here.



left side for negative tweets and right side for positive tweets.

In further analysis we analyzed the tweets based upon emotions.As seen in graph both positive and negative sentiment tweets have highest frequency for happy tweets when compared among themselves.Hence we concluded that whether positive sentiment tweet or negative sentiment tweet both can have any emotion.However every on comparing both graphs positive sentiment graph(right side) has maximum number of emotions happy and negative sentiment graph(left side) has other emotions like sad,surprise at higher side as well. Hence concluding probability of happier emotion in positive tweets is higher.Probability of sad,surprised,anger,fear emotion is higher in negative tweets.



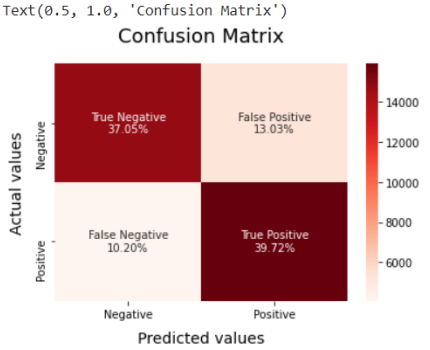
After understanding how emotions and sentiment are correlated,top 10 hashtags for every emotion was found and plotted.Many hashtags are common in all cases , however their frequency vary.These hashtags can also be taken as features for building models.

3) Experimental Results: Testing accuracy when tested upon 20 percent dataset for various models:
Accuracy of LR Model with Cross Validation is: 77.70
Accuracy of Random Forest Model on testing data : 73.60
Accuracy of Multinomial Naive Bayes Model on testing data : 75.59
Accuracy of SVM Model on testing data : 76.81
Accuracy of Perceptron Model on testing data : 71.04

Confusion matrix and experimental result for Logistic Regression model,random forest model,multimonial naive bayes,linear SVM,perceptron respectively.

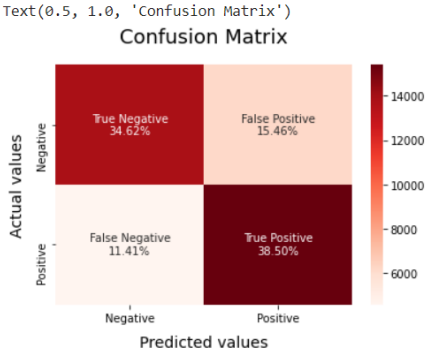
	precision	recall	f1-score	support
0	0.78419	0.73981	0.76135	20035
1	0.75293	0.79569	0.77372	19965
accuracy			0.76770	40000
macro avg	0.76856	0.76775	0.76754	40000
weighted avg	0.76859	0.76770	0.76752	40000

Text(0.5, 1.0, 'Confusion Matrix')



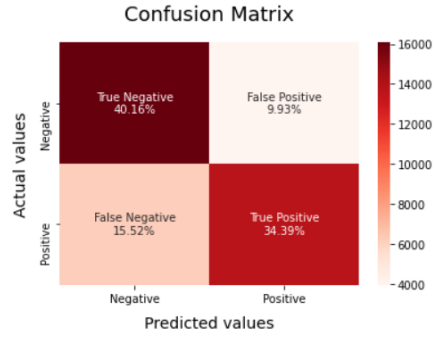
	precision	recall	f1-score	support
0	0.75215	0.69129	0.72043	20035
1	0.71347	0.77140	0.74131	19965
accuracy			0.73128	40000
macro avg	0.73281	0.73135	0.73087	40000
weighted avg	0.73284	0.73128	0.73085	40000

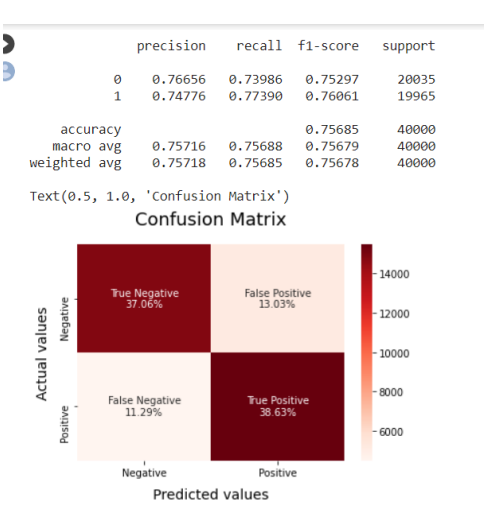
Text(0.5, 1.0, 'Confusion Matrix')

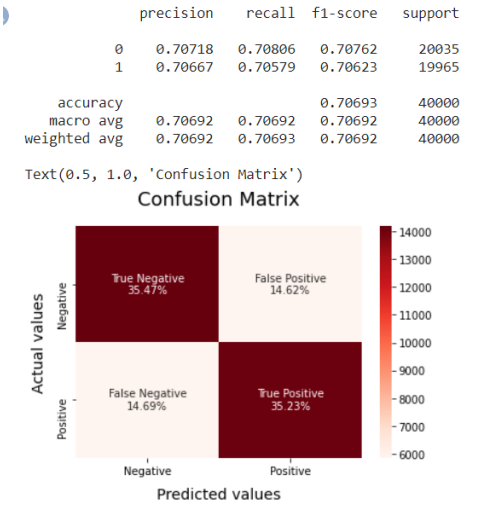


	precision	recall	f1-score	support
0	0.72124	0.80185	0.75941	20035
1	0.77604	0.68901	0.72994	19965
accuracy			0.74552	40000
macro avg	0.74864	0.74543	0.74467	40000
weighted avg	0.74859	0.74552	0.74470	40000

Text(0.5, 1.0, 'Confusion Matrix')







4) Contribution: ALL had contributed to code and analysis.

II. Conclusion

Based on finding and experimental result we conclude that Logistic regression is the best model among all the models used. Perceptron is the most inappropriate model for this dataset.
Emotions can be used to do sentiment analysis.

Appendix A

Appendix B

Acknowledgment

The authors would like to thank their professor Dr. Richa Singh and all the teaching assistants of the course for being extremely supportive during their project and for their guidance through out the course.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

PLACE PHOTO HERE

Pragati Sinha B19CSE065, Department of computer Science and engineering

Khushboo Singh B19EE046, Department of Electrical and engineering

Chirag Bhawnani B19EE022, Department of Electrical and engineering