

Malignant Comments Classifier Project

Submitted by:

PALLAVI BAJPAI

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and references. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

**INTRODUCTION**

* Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

* Conceptual Background of the Domain Problem

A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

* Review of Literature

Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

* Motivation for the Problem Undertaken

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

**Analytical Problem Framing**

* Data Sources and their formats

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels.

* Data Preprocessing Done
* Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 8 columns.
* Duplicate data was removed.
* Outliers were removed but to an extent to maintain 7-8% of data loss.
* Data was imbalanced, so resampling was done.
* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

* State the set of assumptions (if any) related to the problem under consideration

None

* Hardware and Software Requirements and Tools Used
* Python code was written in Jupiter notebook. Below are the libraries needed in the process.
* Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model\_selection.cross\_val\_score, sklearn.linear\_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion\_matrix,accuracy\_score, sklearn.metrics.f1\_score, scipy.stats, seabor.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.

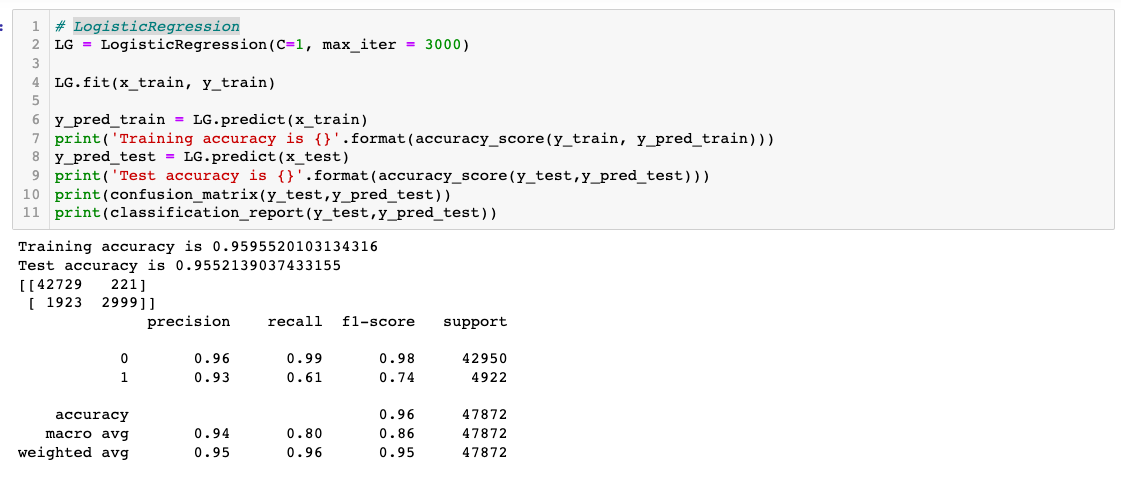
Skewness was checked using histogram.

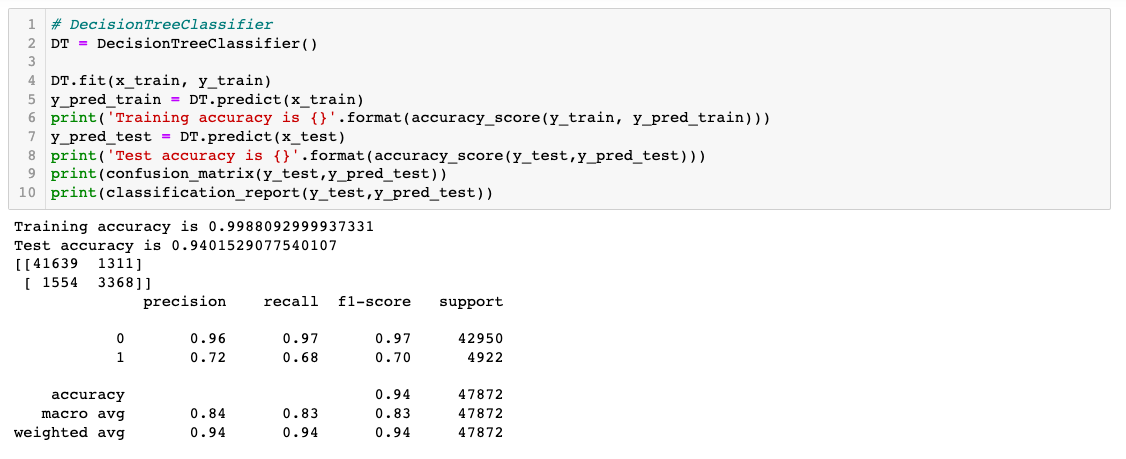
* Testing of Identified Approaches (Algorithms)

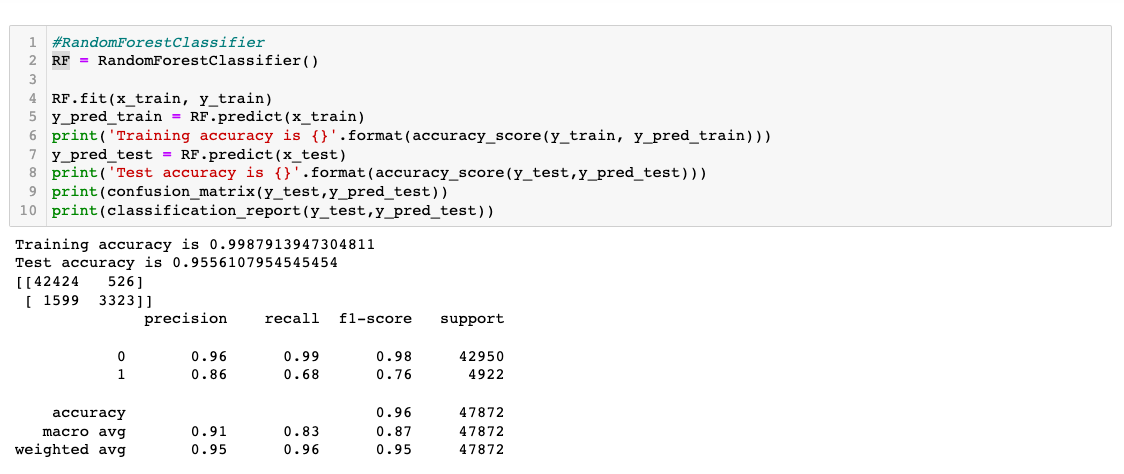
Below algorithms were used for training and testing of data

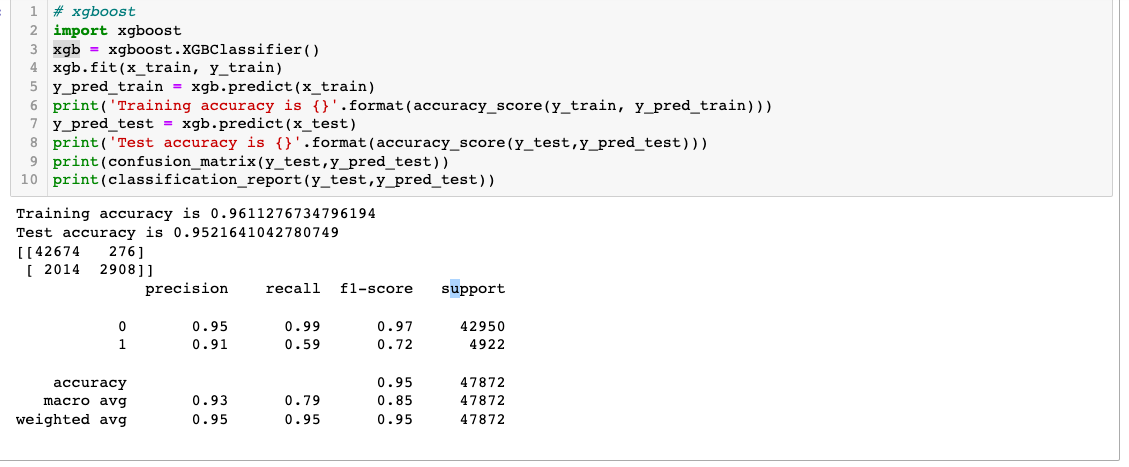
* Random Forest Classifier
* KNeighbors Classifier
* Decision Tree Classifier
* Logistic Regression
* XGboost
* AdaBoost Classifier

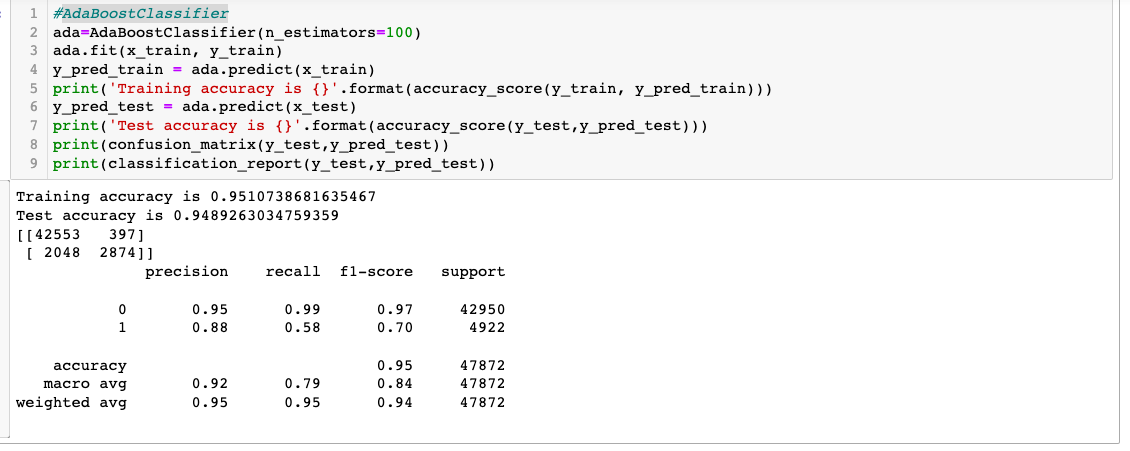
* Run and Evaluate selected models

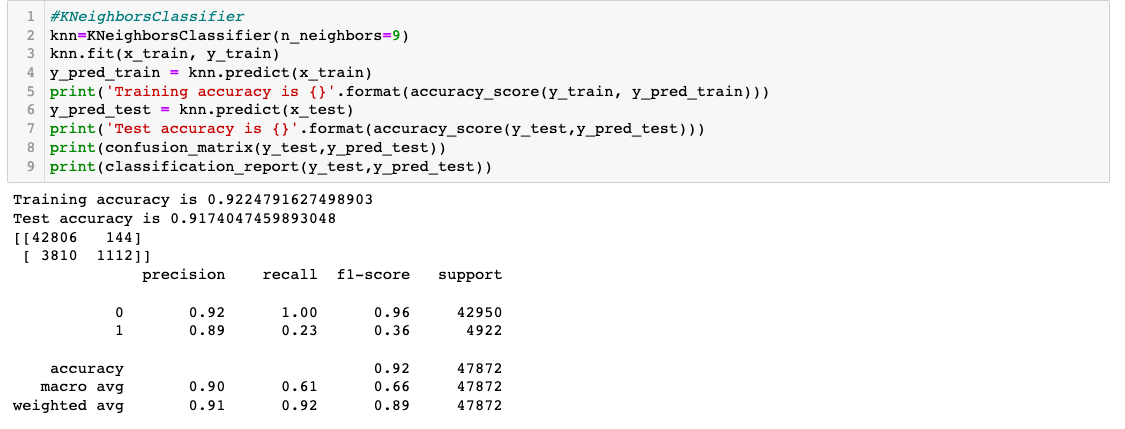




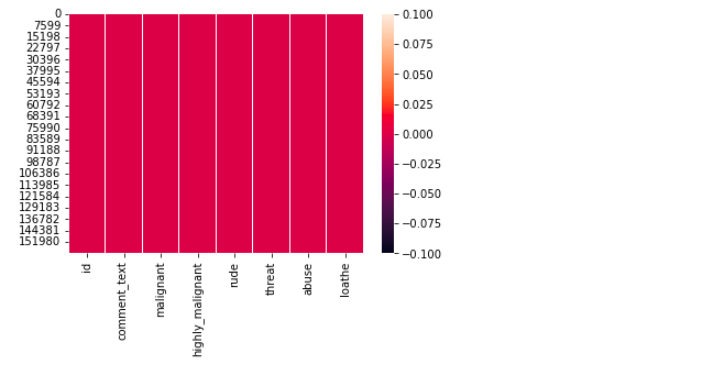


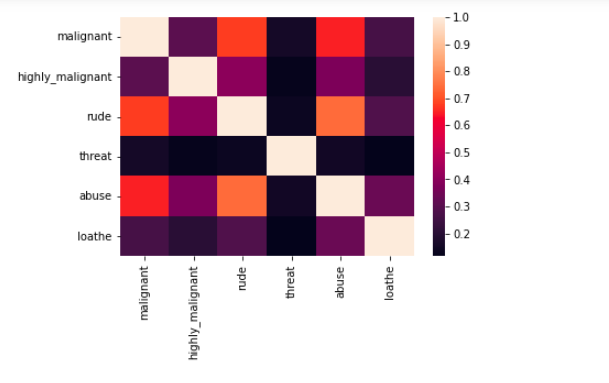


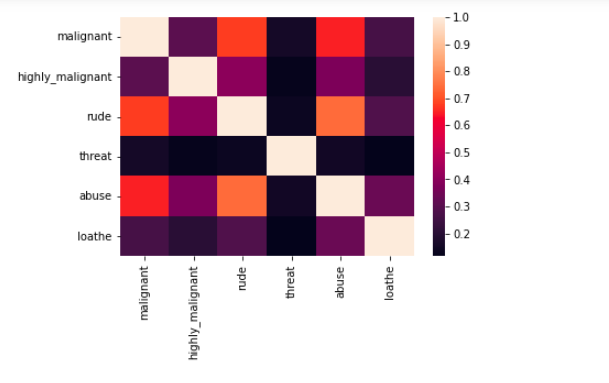
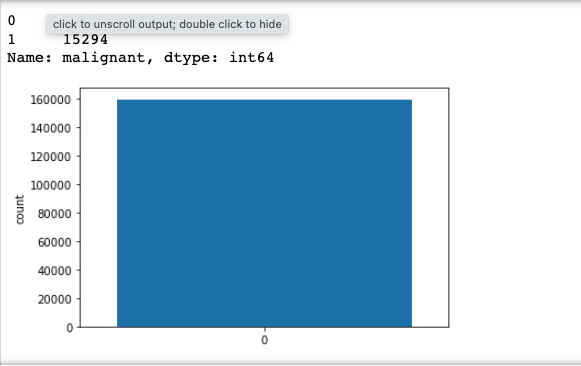


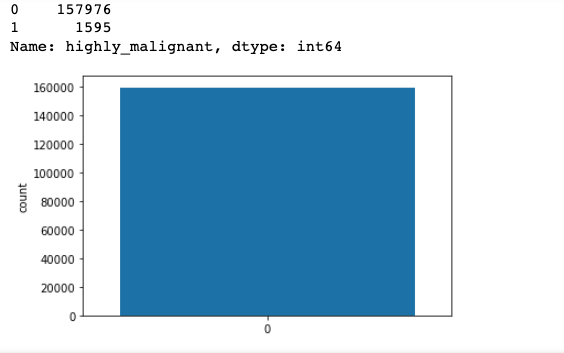


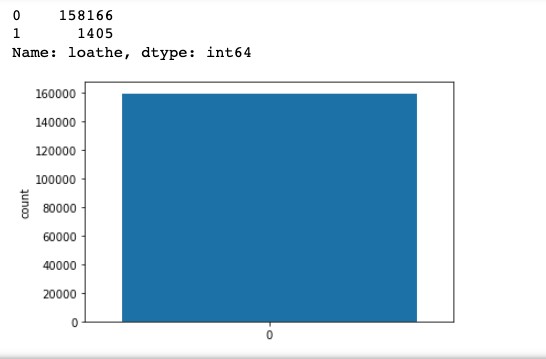
* Key Metrics for success in solving problem under consideration
* Accuracy score was the key metric used to finalize the model.
* Visualizations

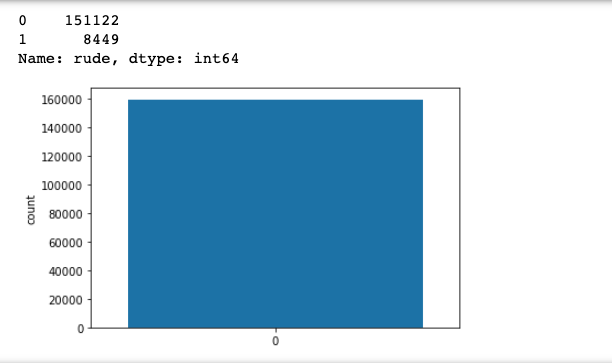


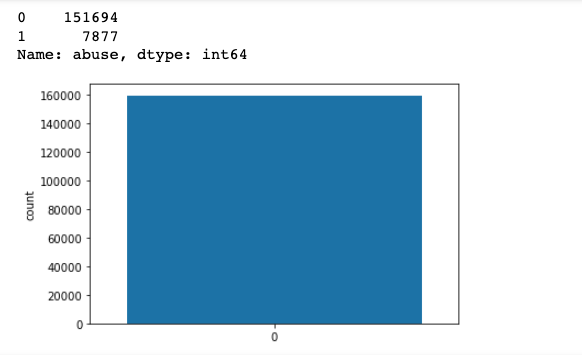


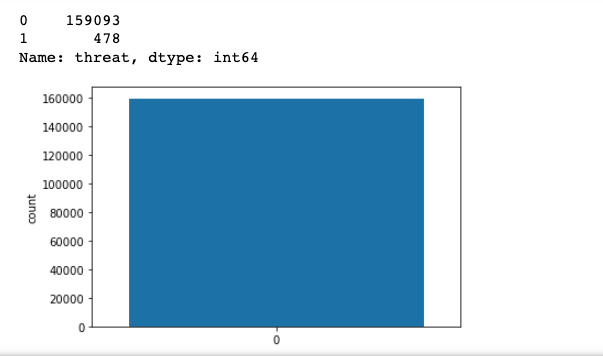
 

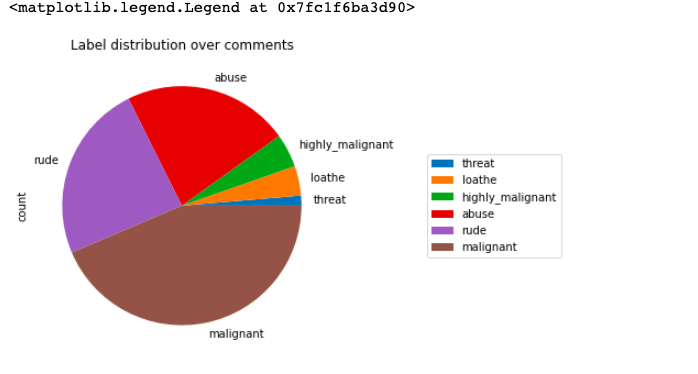


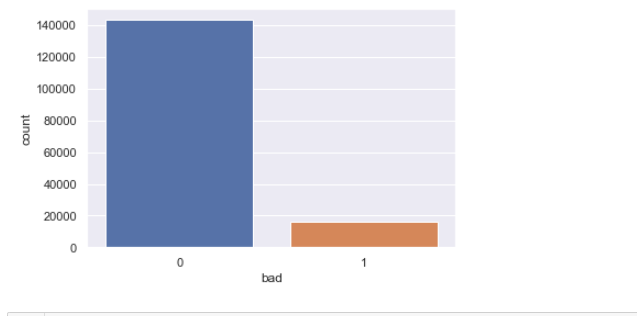


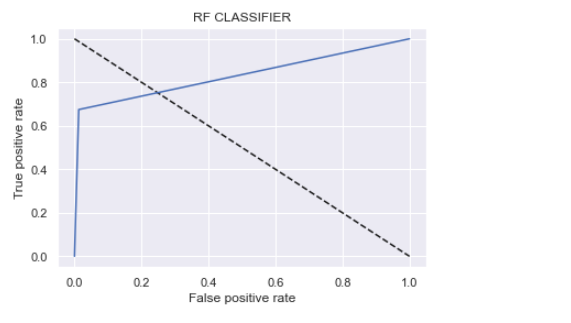




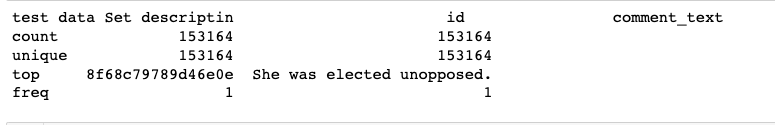








* Interpretation of the Results



* No null values were detected.
* No high correlation.
* Little skewness present
* Most of the comments were Malignant, abuse and rude.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Accuracy score was highest in Random forest Classifier hence we trained the model with the same.

* Learning Outcomes of the Study in respect of Data Science
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.
* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.