

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

Vishnu Rathore

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to “**Flip Robo** **Technologies**” that made this project possible. I would like thank my SME Mr. Keshav Bansal for his guidance in building this project. I would also like to thank Data Trained Institution to making me capable of making proper decisions in the field of damascene and Machine learning. Lastly, I would like to thank my parents to make all of this happen.

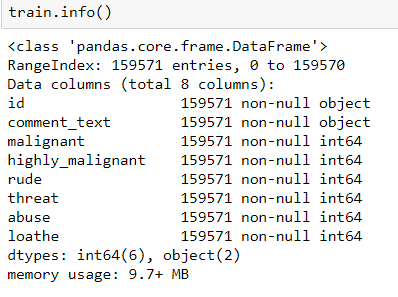
**INTRODUCTION**

* 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 large proportion of online comments present on public domains are usually constructive, however a significant proportion are toxic in nature. Dataset is obtained online which are processed to remove noise from the dataset. The comments contain lot of errors which increases the number of features manifold, making the machine learning model to train the dataset by processing the dataset, in the form of transformation of raw comments before feeding it to the Classification models using a machine learning technique known as the term frequency-inverse document frequency (TF-IDF) technique. The logistic regression technique is used to train the processed dataset, which will differentiate toxic comments from non-toxic comments. The multi-headed model comprises toxicity (severe-toxic, obscene, threat, insult, and identity-hate) or Non-Toxicity Evaluation, using confusion metrics for their prediction.
* So, primarily we must focus which social handles is used mostly among all and what are the moto of particular social handles who is facing this type of comments.
* Review of Literature
* Aggression by text is a complex phenomenon, and different knowledge ﬁelds try to study and tackle this problem.
* This analysis of related workfocuses on a computer science perspective of aggression identiﬁcation, a recent emerging area. Currently, the scientiﬁc study of automatic identiﬁcation of aggressive text, using information technology techniques, is increasing. In this study, several related literature are used to express different types of aggression. Some of those are hate (Tarasova et al.8), cyber bullying (Adamic9), abusive language (Nobata et al.3), toxicity .
* Motivation for the Problem Undertaken
* Because of the most demanding and most exploratory technology i.e. NLP and specifically using machine learning in this type of problem can help me to learn more and explore me more. The following problem statement is also a useful use case because as said earlier comments classification is becoming more important now a days.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Project contain train and test dataset as well.
* In train data set there are 159,571 rows and 8 columns.
* There are no null values in the dataset
* Most of the data are numeric in nature which are binary.
* Comments is object in nature and consist of text.
* Data Sources and their formats

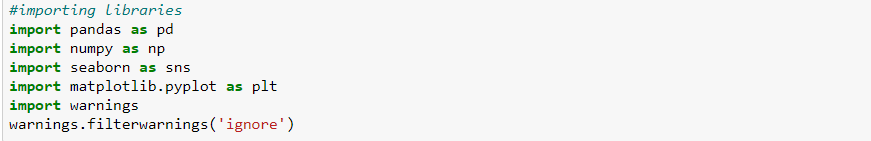
Train Data

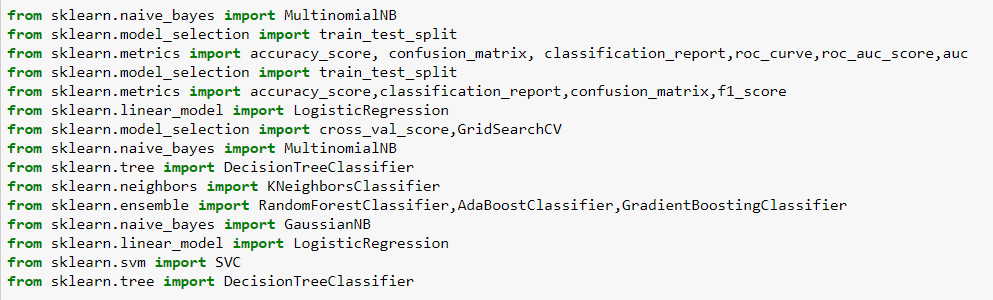


* Data Pre-processing Done



* As there are no null values in the dataset, but as the comment column in text format so there require lot of text pre-processing.
* As, number of offensive comments are in very less in number.
* Data Inputs- Logic- Output Relationships
* There are features which are highly correlated
* It is possible because one comment can be classified into multiple categories
* One comment can be rude and abuse at same time.
* Most of the text length are within 500 characters, with some up to 5,000 characters long
* State the set of assumptions (if any) related to the problem under consideration
* For the easiness of the training and predicting I assumed and make a label of bad words. Means all the bad comments are classify as bad word and all good comments whose value are 0 are classified as good comments
* This will help the model to classify the comment easily also I had made a separate model which helps to predict the prob of comment text that in which column it has high prob.
* Hardware and Software Requirements and Tools Used





* Above are the libraries which I used to pre-process, predict and visualize the project :
* Pandas - It is used to play with data frame and helps in to get more insight of the data, like describing the data and the types of the all features.
* Seaborn, Matplotlib – Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots and so on. Seaborn on the other hand, provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes.
* Sklearn – Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction..

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

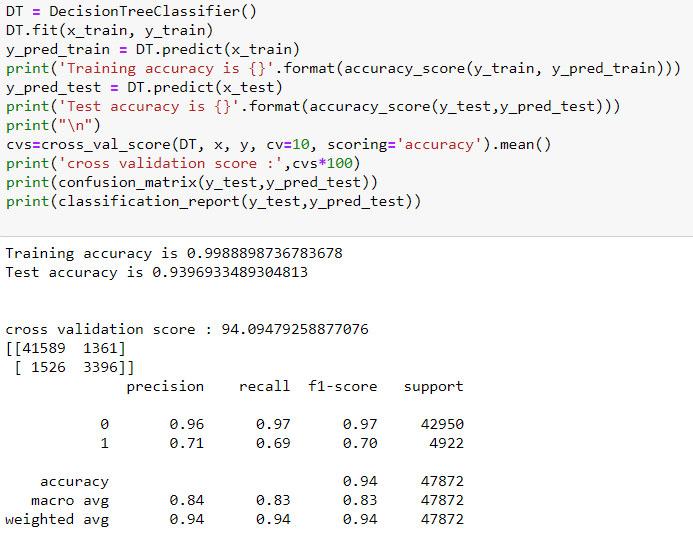
Diagram

Description automatically generated

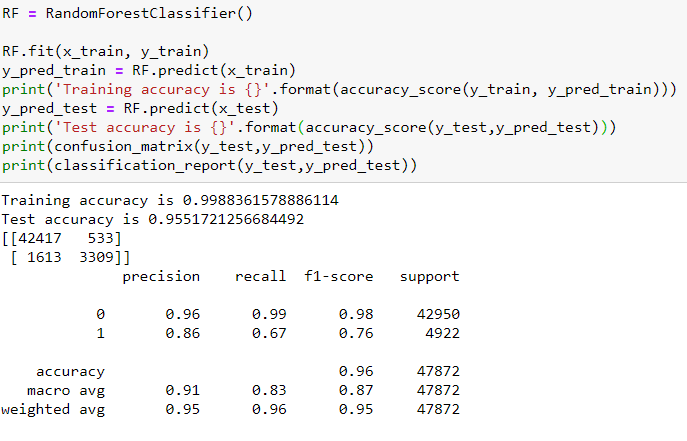
* Read the use case and search from the google to know more about the use case and in which field or domain it will applicable more.
* As data set is already there, so this step is excluded.
* Data Preparation includes the data cleaning and describing the data which I followed text pre-processing.
* EDA involves the visualization which is helpful to get more insight from the data and get to know about the mostly use comment and check the most frequent words.
* Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and apply hyperparameter tuning.
* Model Evaluation, for this I used confusion metrices and mainly focus on False negative and tried to reduce the False negative which is type 2 error and on F1 score as dataset is imbalance.

* Testing of Identified Approaches (Algorithms)

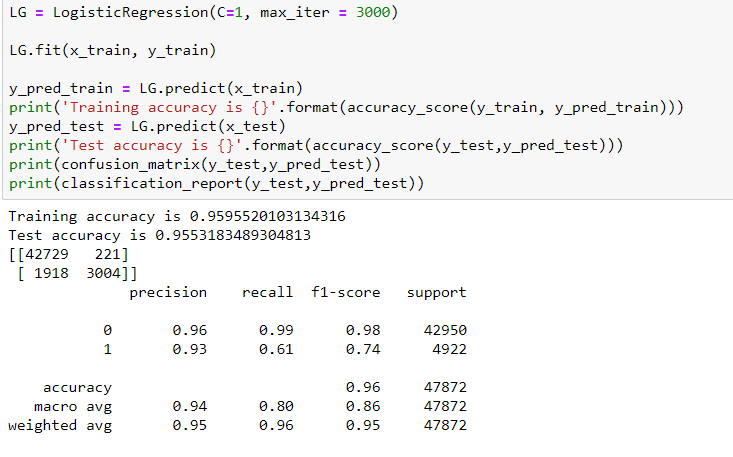
**Decision Tree:**



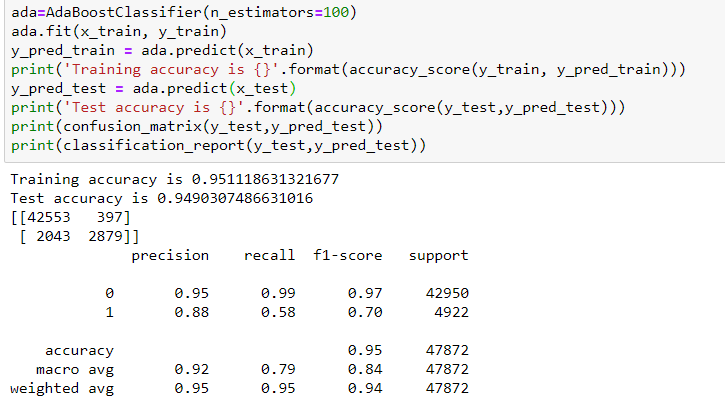
**Random Forest:**



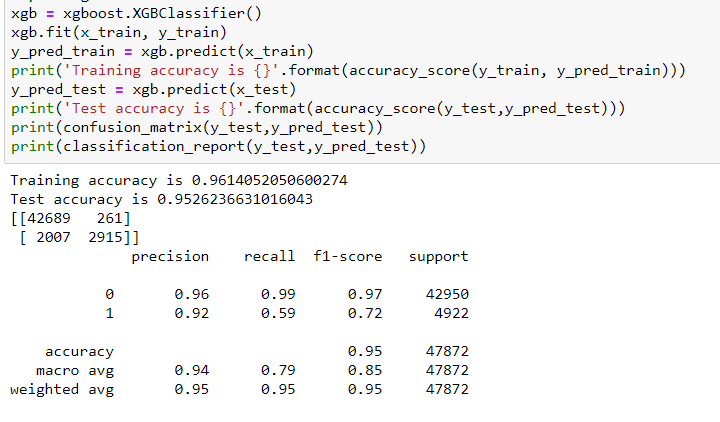
**Logistic regression:**



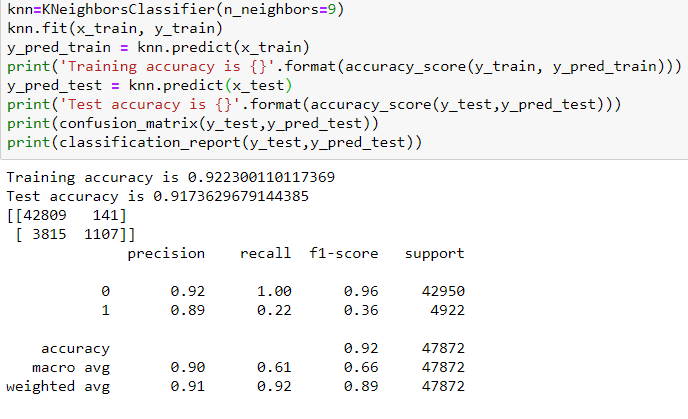
**Ada boost classifier:**



**XG boost classifier:**



**KNN Classifier:**

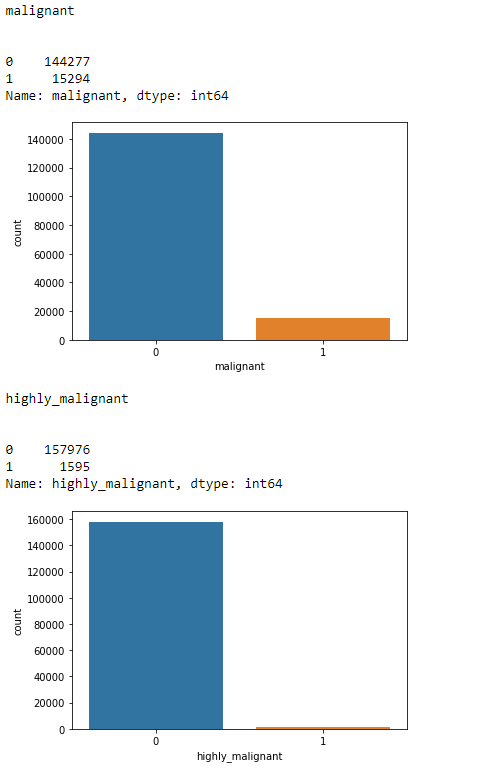


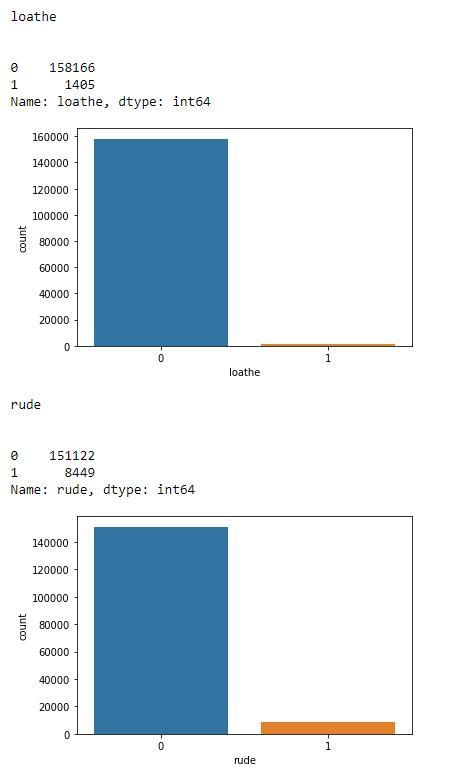
* Run and Evaluate selected models
  + In classification problem there are various metrics that are accuracy score, confusion matrix, classification repot, Roc Auc curve which help to check the efficiency of the model

**Evaluation Matrices:**

* Accuracy - it determines how often a model predicts default and non default correctly.
* Precision-it calculates whenever our models predicts it is default how often it is correct.
* Recall- Recall regulate the actual default that the model is actually predict.
* Precision Recall Curve - PRC will display the trade off between Precision and Recall threshold.
* F1 score - the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.
* Cross Validations:
* K Fold cross validations , K = 10
  + So, in this case accuracy score is good but most important is confusion matrix in which we must decrease the False Positive that is type 2 error.
* Visualizations

Mostly I used seaborn, word cloud and matplotlib for visualization and EDA

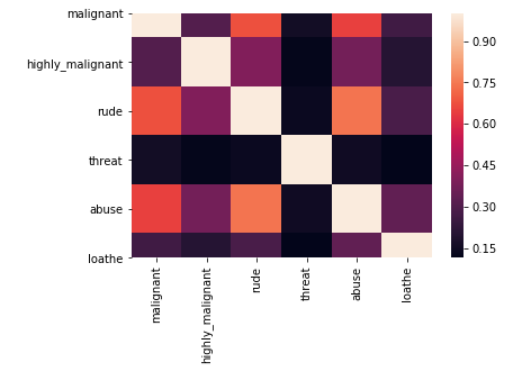




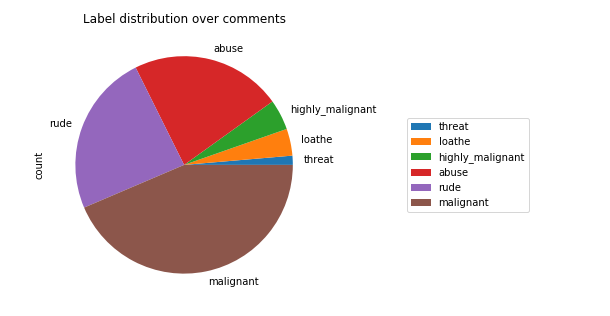
Wordcloud:



Correlation heatmap :



Abuse and malignant, rude and abuse are highly correlated.



**CONCLUSION**

* Key Findings and Conclusions of the Study
* In this project there are some variables like malignant and rude which are highly correlated it is possible because one comment text may have combination of multiple features.
* Removing the column id does not impact the model training.
* Using Tree, model can reduce the false negative values
* It has future scope in various use cases likewise in election, social media etc, where every day there are multi offensive comments spread.
* So, in the future it may use very well to easily classify the comments as bad or good.
* Random forest is well suitable for this project as it used tree internally and it used multiple weak learner and generate the strong model and generate low bias and low variance model.
* Learning Outcomes of the Study in respect of Data Science
* While implementing this project most of the time is taken doing the evaluation of metrices.
* Gain knowledge and get more insight of various stemmer and vectorizer.
* Tree algorithms are well suitable for this project as it used tree internally and it used multiple weak learner and generate the strong model and generate low bias and low variance model.
* Limitations of this work and Scope for Future Work
* It has future scope in various use cases likewise in election, social media etc, where every day there are multi offensive comments spread.
* So, in the future it may use very well to easily classify the comments as bad or good.