**NAME OF THE PROJECT**

**Malignant Commentes Classifier - Multi Label Classification Project  
Submitted by:**

**SAUNAK MUKHERJEE**



I would like to express my special gratitude to THE “Flip Robo” team, they have given me this opportunity to deal with a beautiful dataset and it has

helped me to improve analyzation of skills. And I want to express my

huge gratitude to Ms. Sapna Verma (SME Flip Robo), she is the person

who has helped me When I am in difficulties I faced while doing the

project. And also A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo.

References use in this project:

1. SCIKIT Learn Library Documentation

2. Blogs from towardsdatascience, Analytics Vidya, Medium

3. Andrew Ng Notes on Machine Learning (GitHub)

4. Data Science Projects with Python Second Edition by Packt

5. Hands on Machine learning with scikit learn and tensor flow by

Aurelien Geron

6. Stackoverflow.com to resolve some project related queries.

7. Predicting Credit Default among Micro Borrowers in Ghana

Kwame Simpe Ofori, Eli Fianu

8. Predicting Microfinance Credit Default: A Study of Nsoatreman

Rural Bank, Ghana Ernest Yeboah Boateng

9. A Machine Learning Approach for Micro-Credit Scoring

Apostolos Ampountolas

And also thank you for many other persons who has helped me directly

or indirectly to complete the project

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. 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. **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.**

**The data set includes:**

* Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
* Highly Malignant: It denotes comments that are highly malignant and hurtful.
* Rude: It denotes comments that are very rude and offensive.
* Threat: It contains indication of the comments that are giving any threat to someone. Abuse: It is for comments that are abusive in nature.
* Loathe: It describes the comments which are hateful and loathing in nature.
* ID: It includes unique Ids associated with each comment text given.
* Comment text: This column contains the comments extracted from various social media platforms.

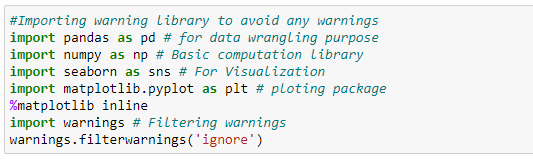
**THE OVERVIEW OF THE MODEL**

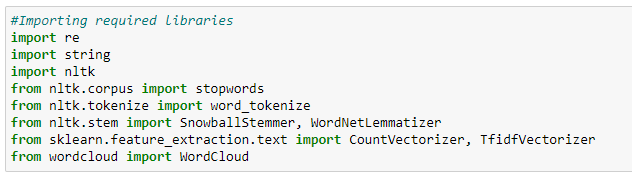
The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’. The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment. The data set includes:Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not. Highly Malignant: It denotes comments that are highly malignant and hurtful. Rude: It denotes comments that are very rude and offensive.Threat: It contains indication of the comments that are giving any threat to someone. Abuse: It is for comments that are abusive in nature. Loathe: It describes the comments which are hateful and loathing in nature. ID: It includes unique Ids associated with each comment text given. Comment text: This column contains the comments extracted from various social media platforms. This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes many categories of comments, we can do good amount of data exploration and derive some interesting features using the comments text column available. You need to build a model that can differentiate between comments and its categories. Refer to the data set file provided along with this.

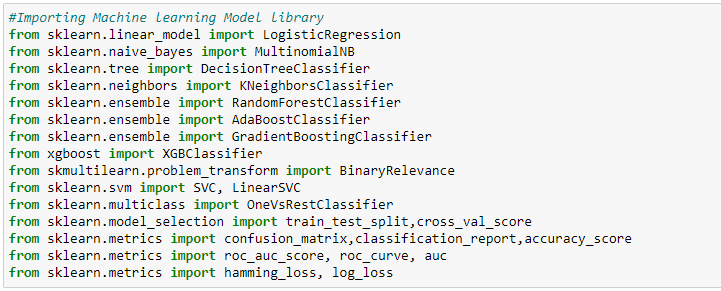
Data Sources and their formats

The data set comes from my internship company – Fliprobo technologies

in excel format



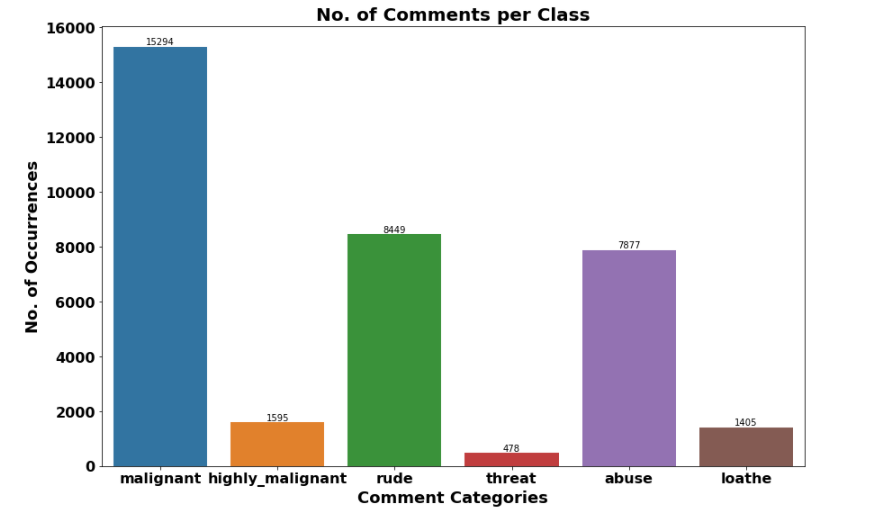




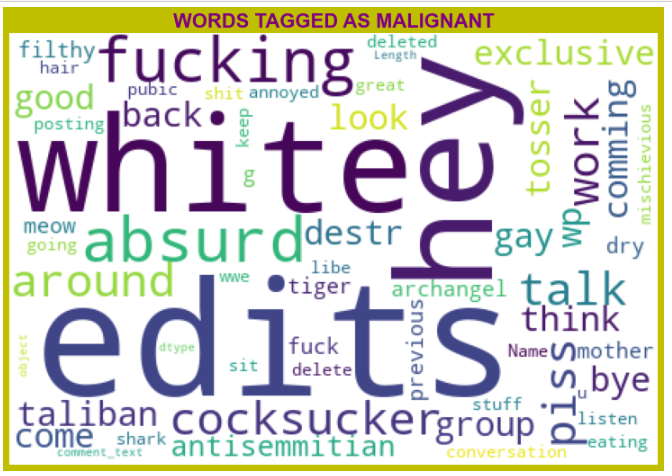
**Machine Learning Model Building**

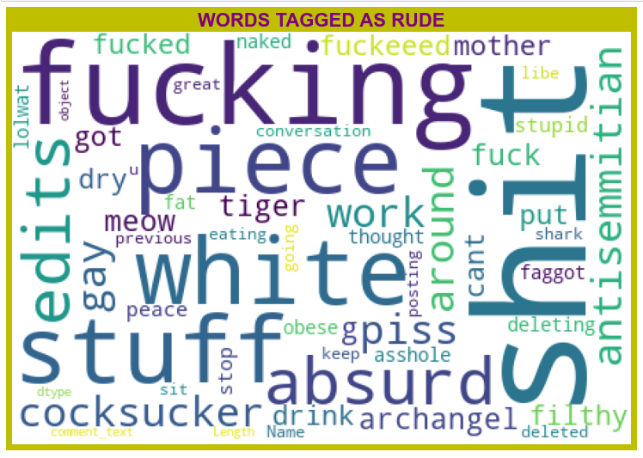
The different classification algorithm used in this project to build ML model are as below:Random Forest classifierSupport Vector ClassifierLogistics RegressionAdaBoost Classifier

**Target Variable Ratings**

We have seen around 90% comments are good and 10% are nagetive.

**Word Cloud**

Word Cloud which is the visualization technique for text data where we can use each word that is picturized with its importance of the context or its frequency.The more commonly which is the term of appears within the text being with analysed, the larger the word appears in the image which is generated.The enlarged texts that are the greatest number of the words used for the and small texts which are the smaller number of the words used.

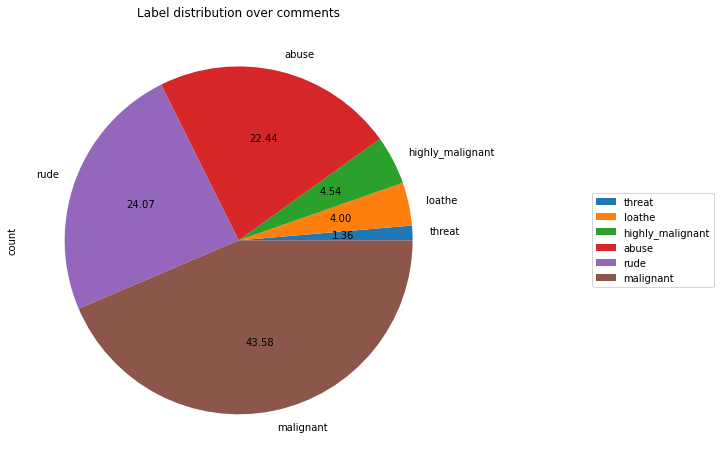


From word cloud of Rude comments, it is clear that it mostly consists of words like fucking, shit, white, piece, stupid,die,gay, stuff, absurd etc

***Machine Learning Model Building***

We are using this classifier Random Forest classifier,Support Vector Classifier,Logistics Regression,AdaBoost Classifier.**Hyper parameter Tuning**

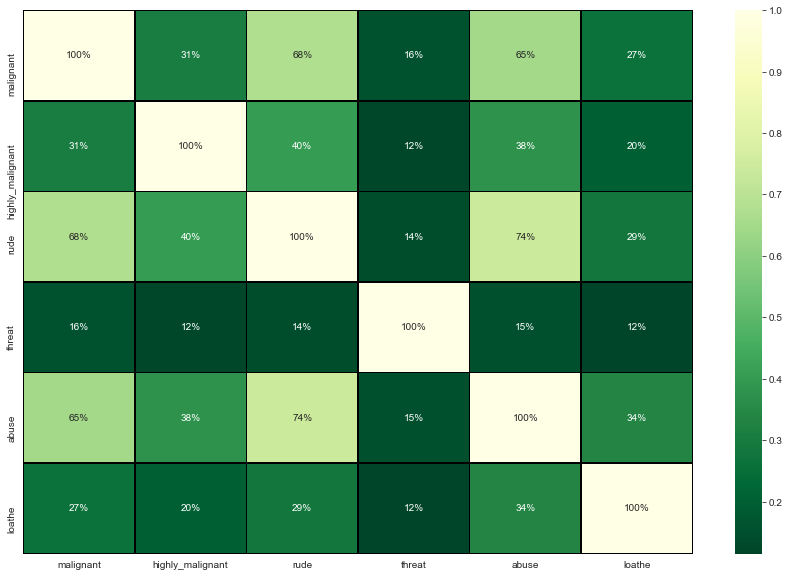
Hyper parameter Tuning is perform over this best model using best param shown below :



 Around 90% comments are Good/Neutral in nature while rest 10% comments are Negative in nature.

 Out of total negative comments around 43.58% are malignant in nature followed by 24.07% are rude comments**.**

**CORRELATION**

****

* **The highest positive correlation is seen in between fields 'rude' and 'abuse'.**
* **Attribute 'threat' is negatively correlated with each and every other feature of this training dataset.**

**Almost all variable are correlated with each other negatively.\***

**CONCLUSION**

Hyperparameter Tuning which is giving us Accuracy score of 91.26% which is slightly improved compare to earlier Accuracy score of 91.15%.