

**Malignant Comments Classifier Prediction Project**

**Submitted by:**

**Devanshu Suri**

**ACKNOWLEDGMENT**

I would like to express my very great appreciation to my SME Ms. Astha Mishra for his valuable and constructive suggestions during the planning and development of this research work. Her willingness to give his time so generously has been very much appreciated.

Separately, I would like to thank:

* FlipRobo Technologies team
* Data Trained Team

Research papers that helped me in this project was as follows:

* <https://medium.com/@dobko_m/nlp-text-data-cleaning-and-preprocessing-ea3ffe0406c1>
* https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

Articles that helped me in this project was as follows:

* [TF-IDF Vectorizer scikit-learn. Deep understanding TfidfVectorizer by… | by Mukesh Chaudhary | Medium](https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a)

**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.
* 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.
* **Conceptual Background of the Domain Problem**
* In the past few years its seen that the cases related to social media hatred have increased exponentially. The social media is turning into a dark venomous pit for people now a days. Online hate is the result of difference in opinion, race, religion, occupation, nationality etc.
* In social media the people spreading or involved in such kind of activities uses filthy languages, aggression, images etc. to offend and gravely hurt the person on the other side. This is one of the major concerns now.
* The result of such activities can be dangerous. It gives mental trauma to the victims making their lives miserable. People who are not well aware of mental health online hate or cyber bullying become life threatening for them. Such cases are also at rise. It is also taking its toll on religions. Each and every day we can see an incident of fighting between people of different communities or religions due to offensive social media posts.
* Online hate, described as abusive language, aggression, cyberbullying, hatefulness, insults, personal attacks, provocation, racism, sexism, threats, or toxicity has been identified as a major threat on online social media platforms. These kinds of activities must be checked for a better future.
* **Motivation for the Problem Undertaken**
* The project was the first provided to me by FlipRobo as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary objective. However, the motivation for taking this project was that it is relatively a new field of research. Here we have many options but less concrete solutions. The main motivation 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**

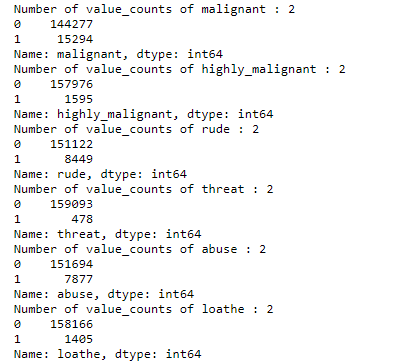
* **Data Sources and their formats**
* The data was provided by FlipRobo in CSV format. After loading the training dataset into Jupyter Notebook using Pandas and it can be seen that there are eight columns named as:

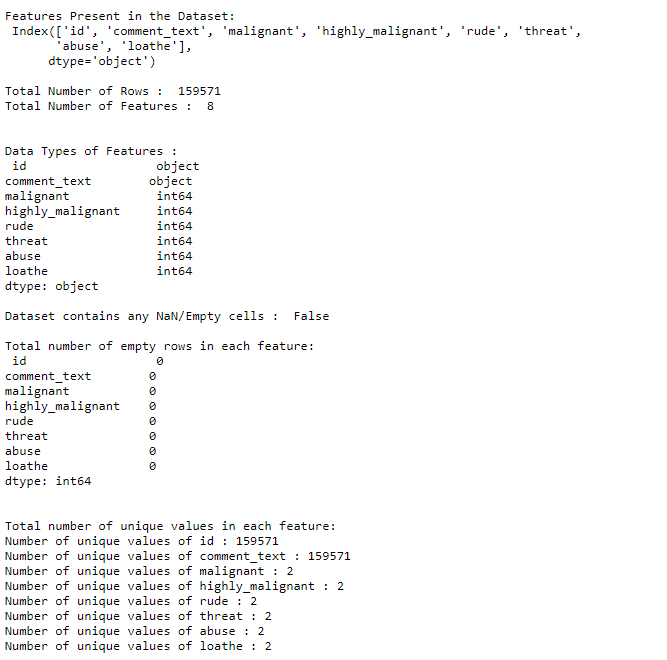
**“** id, comment\_text, “malignant, highly\_malignant, rude, threat, abuse, loathe**”.**

There are 8 columns in the dataset provided:

The description of each of the column is given below:

* **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.





* **Data Pre-processing Done**

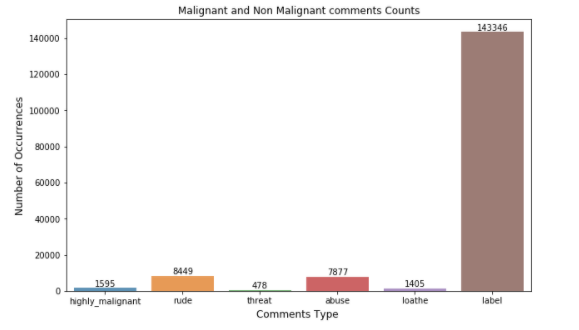
****

For Data pre-processing we did some data cleaning, where we used wordNet lemmatizer to clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizers and joined and return the filtered words.

Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms.

* **Data Inputs- Logic- Output Relationships**

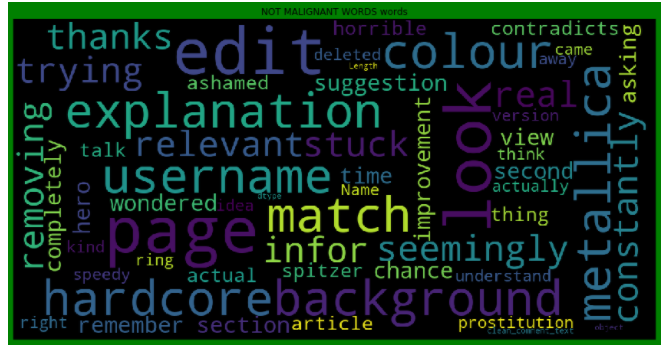
**Comments categories counts:**

****

**Malignant Words:**

****

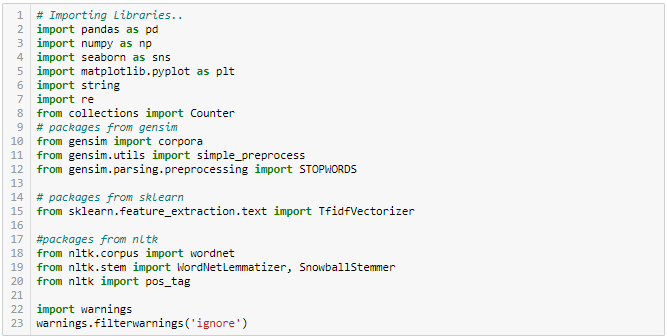
**Not Malignant Words:**

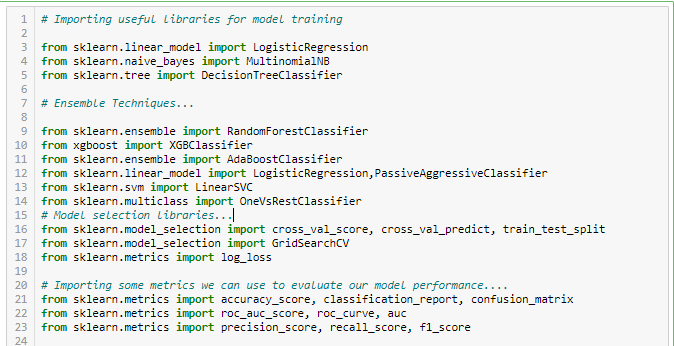
****

From the above we can see that most frequent words for both Malignant and Non Malignant category.

* **Hardware and Software Requirements and Tools Used**
  + Hardware: 8GB RAM, 64-bit, 9th gen i7 processor.
  + Software: MS-Excel, Jupyter Notebook, python 3.6.

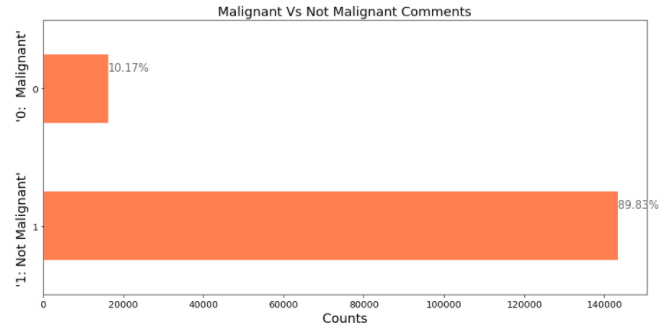
**Libraries used:-**

****

****

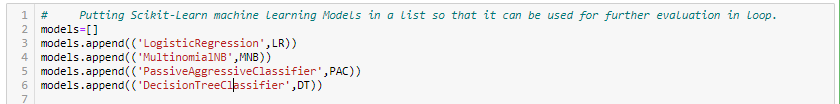
**Model/s Development and Evaluation**

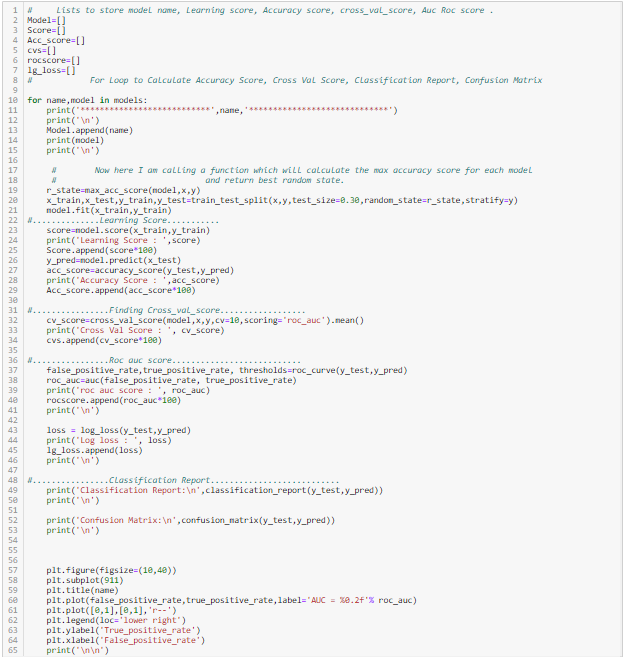
* **Identification of possible problem-solving approaches (methods).**

****

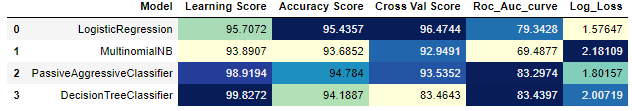
* **Testing of Identified Approaches (Algorithms)**
* LR=LogisticRegression()
* MNB=MultinomialNB()
* PAC=PassiveAggressiveClassifier()
* DT=DecisionTreeClassifier()

**Run and Evaluated selected models**

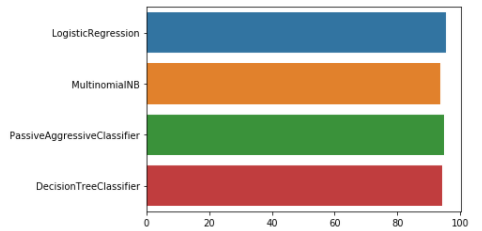




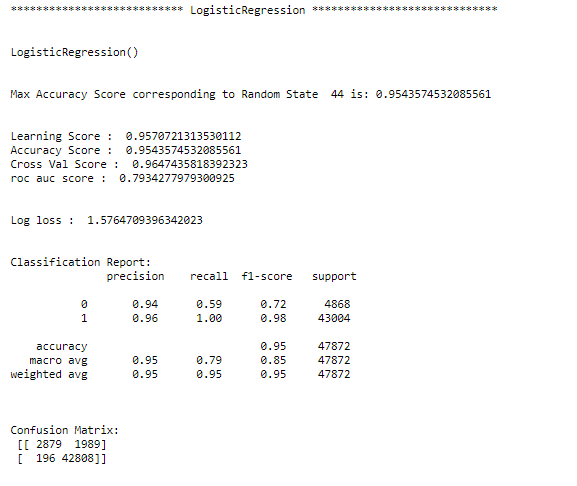
* **Key Metrics for success in solving problem under consideration**

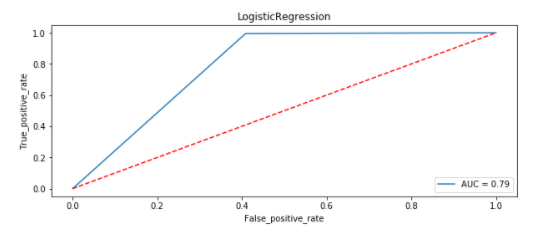


Key Metrices used were the Accuracy Score, Crossvalidation Score and AUC & ROC Curve as this was binary classification. From the above we can see that there are various models out of which we few gave good accuracy score as more than 90%,

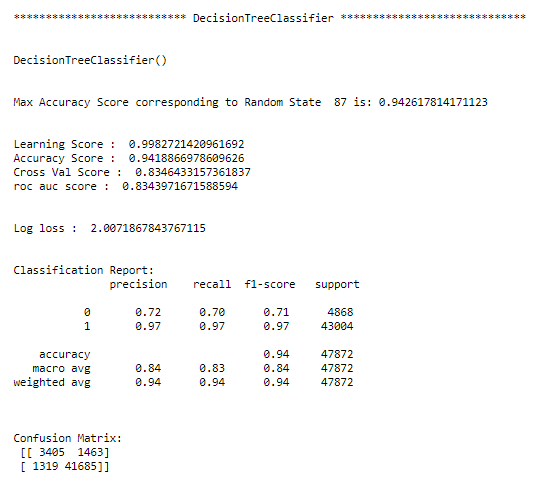


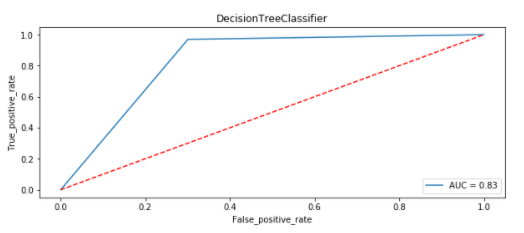
* **Visualizations:**
* Logistic regression:



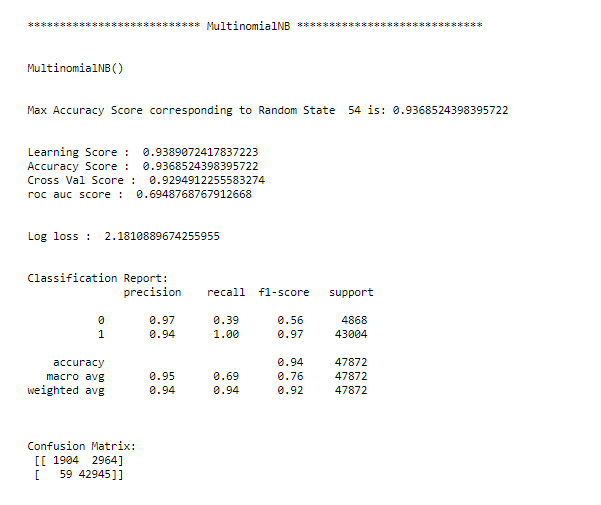


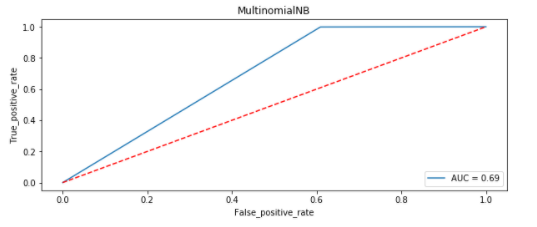
* Decision Tree Classifier:

****

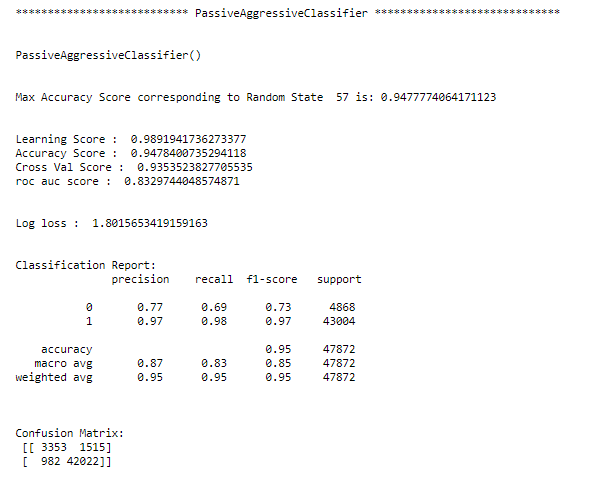
****

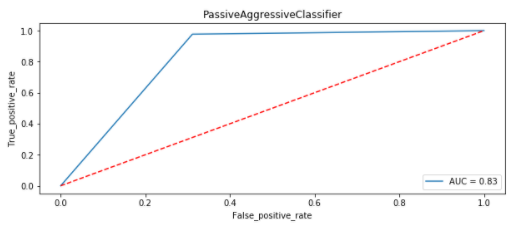
* MultiNomial NB:

****

****

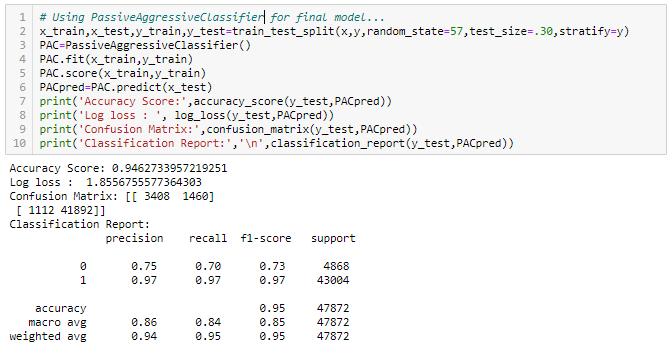
* Passive Aggressive Classifier:

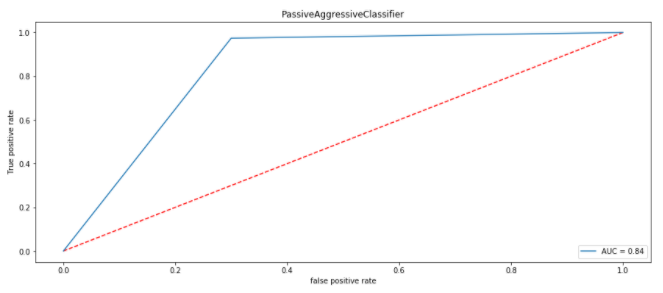
****

****

# After all this process conclusion is that Passive Aggressive Classifier is giving accuracy of 94.78%. So, now I am making a final model using Passive Aggressive Classifier.

* **Final Model:**

****



# After all this process conclusion is that Passive Aggressive Classifier is giving accuracy of 94.78%, and with AUC\_ROC score 0f 84%. So, now I am making a final model using Passive Aggressive Classifier.

* **Interpretation of the Results**
* From the above visualization and matrices found that the Passive Aggressive Classifier performed the best AUC\_ROC\_SCORE **i.e. 94.78%.**

**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* 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.
* From the above analysis the below mentioned results were achieved which depicts the

chances and conditions of a comment being a hateful comment or a normal comment.

* With the increasing popularity of social media, more and more people consume feeds from social media and due differences they spread hate comments to instead of love and harmony. It has strong negative impacts on individual users and broader society.
* **Learning Outcomes of the Study in respect of Data Science**

It is possible to classify the comments content into the required categories of Malignant and Non Malignant. However, using this kind of project an awareness can be created to know what is good and bad. It will help to stop spreading hatred among people.

* **Limitations of this work and Scope for Future Work**
* Machine Learning Algorithms like Decision Tree Classifier took enormous amount of time to build the model and Ensemble techniques were taking a lot more time thus I have not included Ensemble models.
* Using Hyper-parameter tuning would have resulted in some more accuracy.
* Every effort has been put on it for perfection but nothing is perfect and this project is of no exception. There are certain areas which can be enhanced. Comment detection is an emerging research area with few public datasets. So, a lot of works need to be done on this field.