

***MALIGNANT COMMENTS CLASSIFIER PROJECT***

**Submitted by:**

**Shubham Sahu**

# ACKNOWLEDGMENT

I would like to express my very great appreciation to my SME Ms. Khushboo Garg for her 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 were 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 were as follows:

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

*TABLE OF CONTENTS*

[**ACKNOWLEDGMENT** 2](#_Toc73948670)

[**INTRODUCTION** 1](#_Toc73948671)

[BUSINESS PROBLEM FRAMING 1](#_Toc73948672)

[CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM 1](#_Toc73948673)

[REVIEW OF LITERATURE 2](#_Toc73948674)

[MOTIVATION FOR THE PROBLEM UNDERTAKEN 2](#_Toc73948675)

[**ANALYTICAL PROBLEM FRAMING** 3](#_Toc73948676)

[MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM 3](#_Toc73948677)

[DATA SOURCES AND THEIR FORMATS 3](#_Toc73948678)

[DATA PREPROCESSING DONE 5](#_Toc73948679)

[DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS 9](#_Toc73948680)

[HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED 11](#_Toc73948681)

[**MODEL/S DEVELOPMENT AND EVALUATION** 12](#_Toc73948682)

[IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS) 12](#_Toc73948683)

[TESTING OF IDENTIFIED APPROACHES (ALGORITHMS) 13](#_Toc73948684)

[VISUALIZATIONS 13](#_Toc73948685)

[**CONCLUSION** 31](#_Toc73948686)

[KEY FINDINGS AND CONCLUSIONS OF THE STUDY 31](#_Toc73948687)

[LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE 31](#_Toc73948688)

[LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK 31](#_Toc73948689)

# 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.

## REVIEW OF LITERATURE

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.

## 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

## MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM

Here we are dealing with one main text columns which held some importance of the data and others shows the multiple types of behaviour inferred from the text. I prefer to select on focus more on the words which has great value of importance in the context. Countvector is the NLP terms I am going to apply on text columns. This converts the important words proper vectors with some weights.

## 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:

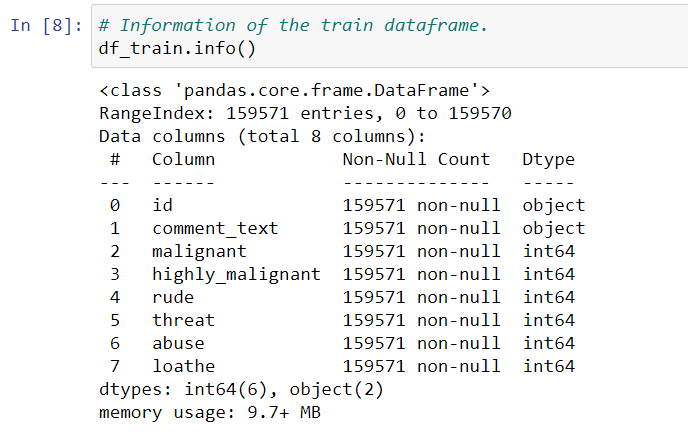
**“** id, comment\_text, “malignant, highly\_malignant, rude, threat, abuse, loathe**”.**

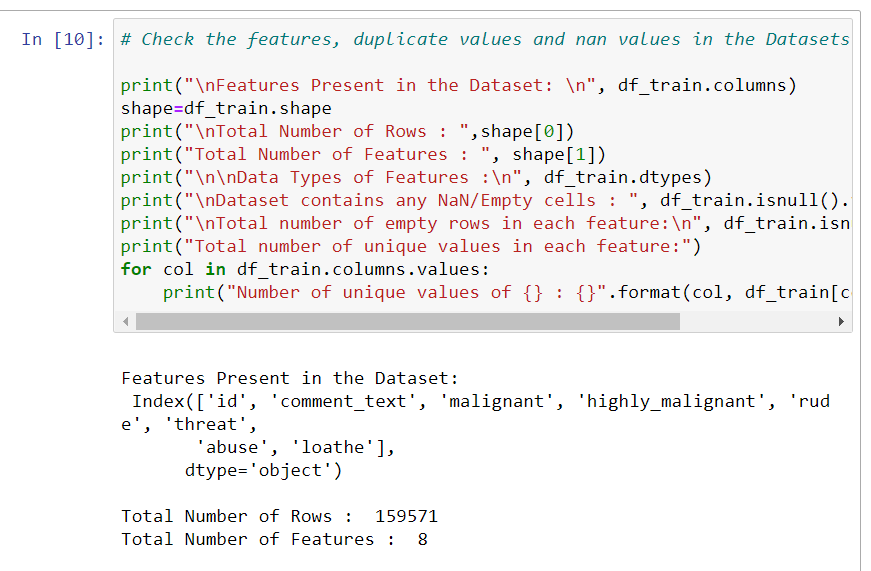
There are 8 columns in the dataset provided:

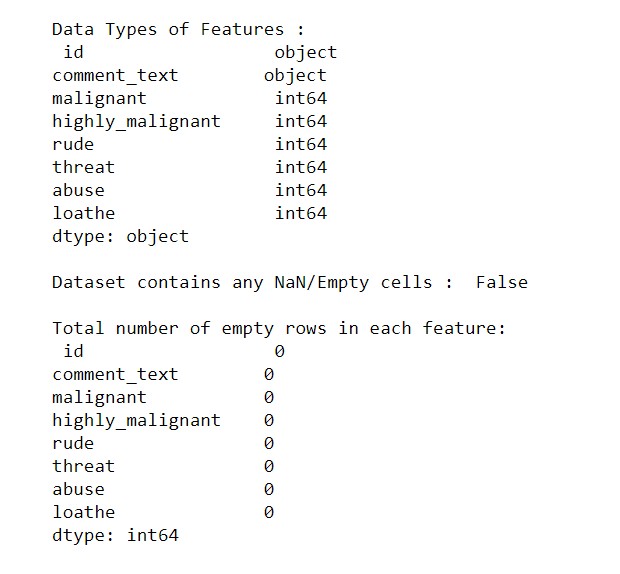
The description of each of the column is given below:

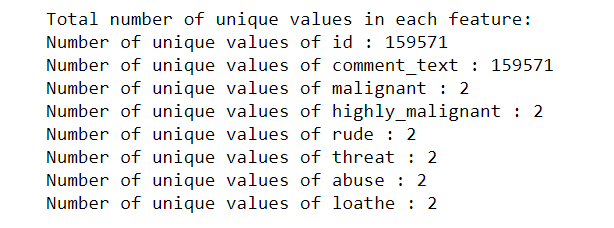
* **Malignant:** It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
* **Highly Malignant:** It denotes comments that are highly malignant and hurtful.
* **Rude:** It denotes comments that are very rude and offensive.
* **Threat:** It contains indication 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.

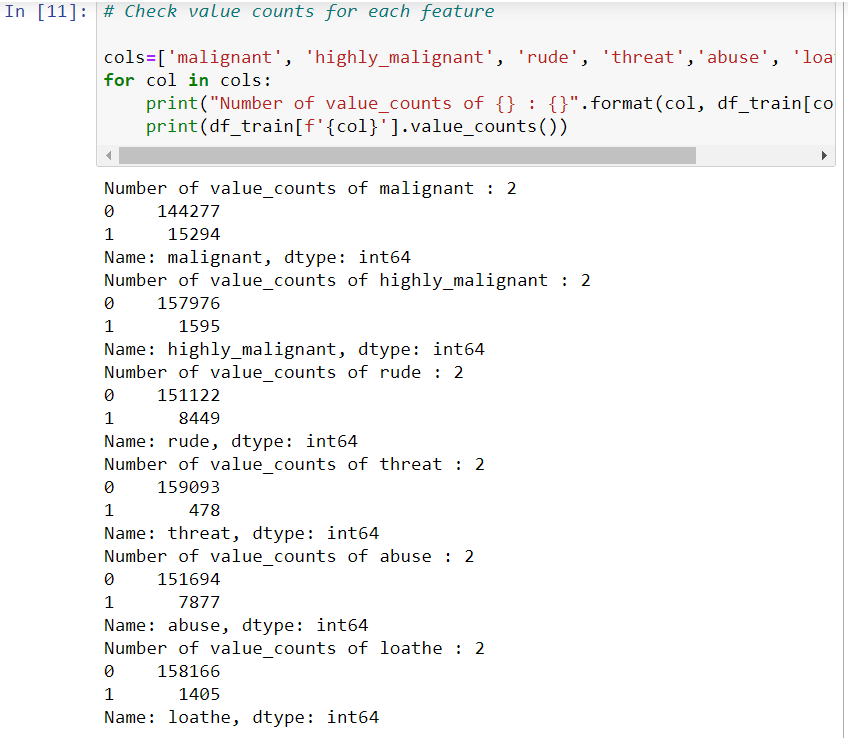
**Comment text:** This column contains the comments extracted from various social media platforms.





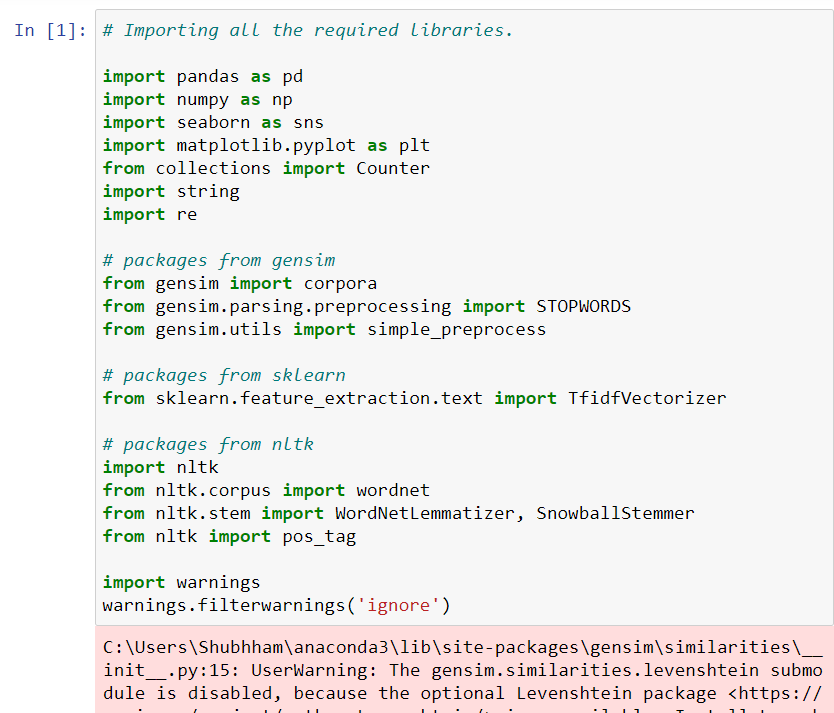




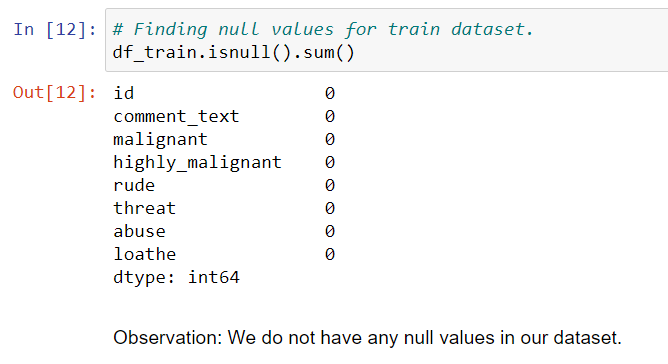


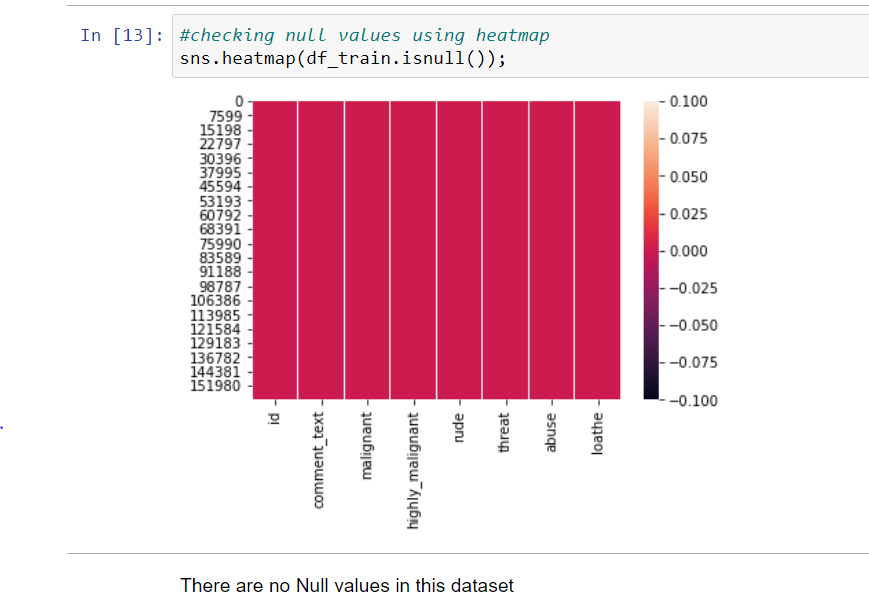
## DATA PREPROCESSING DONE

After loading all the required libraries we loaded the data into our jupyter notebook.



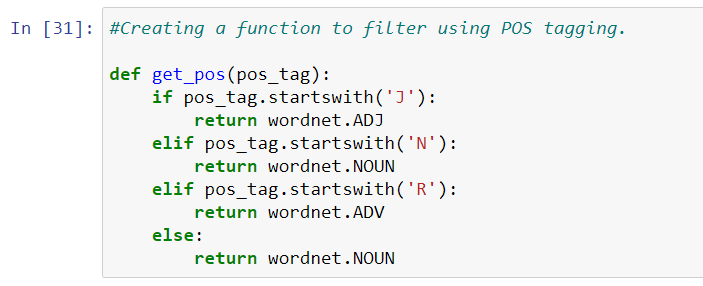
Feature Engineering has been used for cleaning of the data. We first did data cleaning. We first looked percentage of values missing in columns.

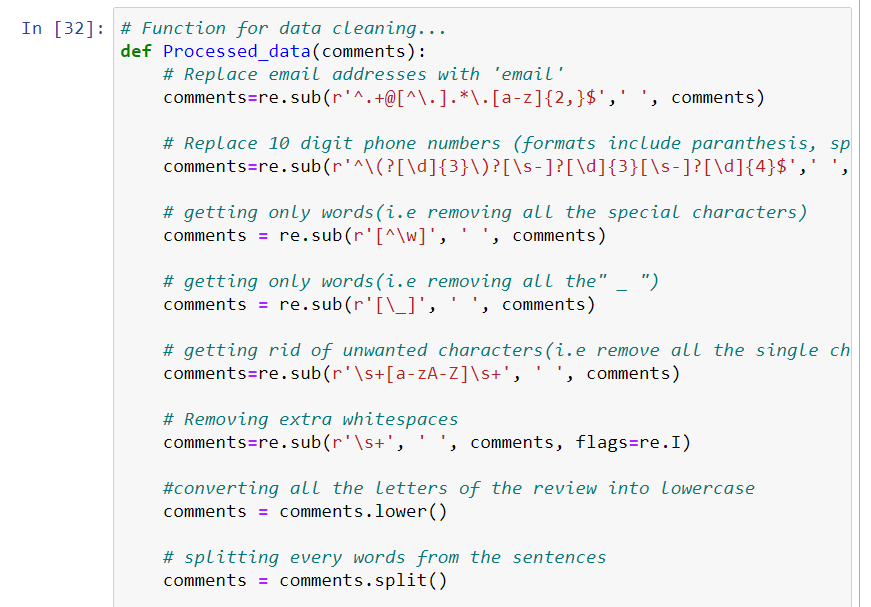


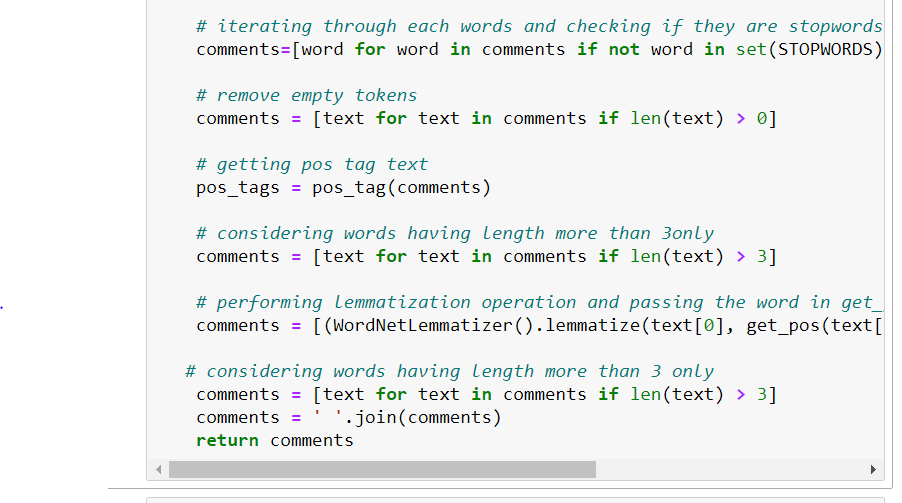


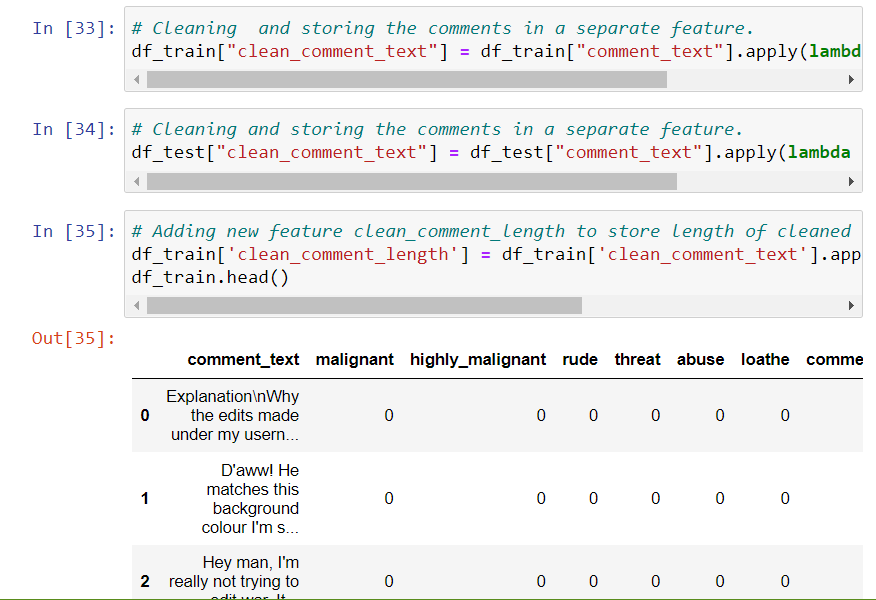
For Data pre-processing we did some data cleaning, where we used wordNetlemmatizerto clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizers and joined and return the filtered words.

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



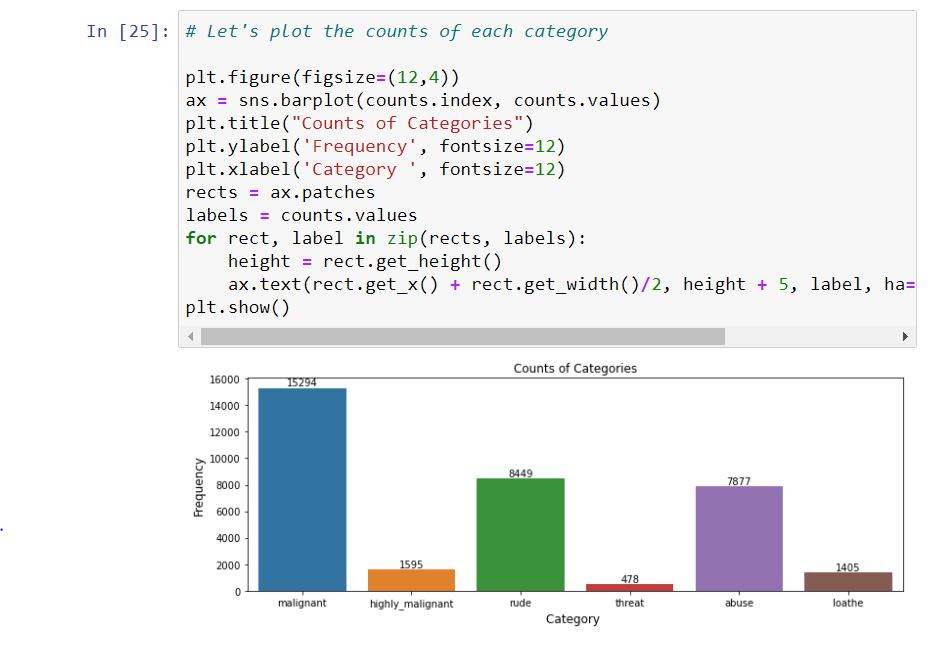




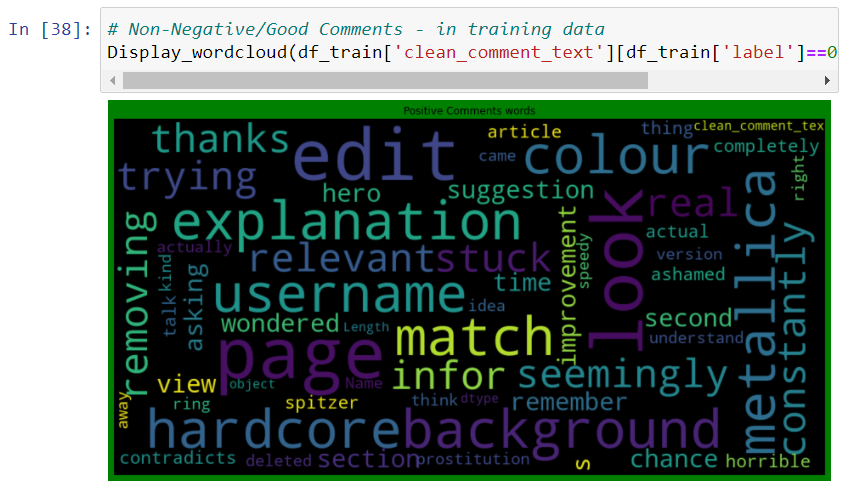


## DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS

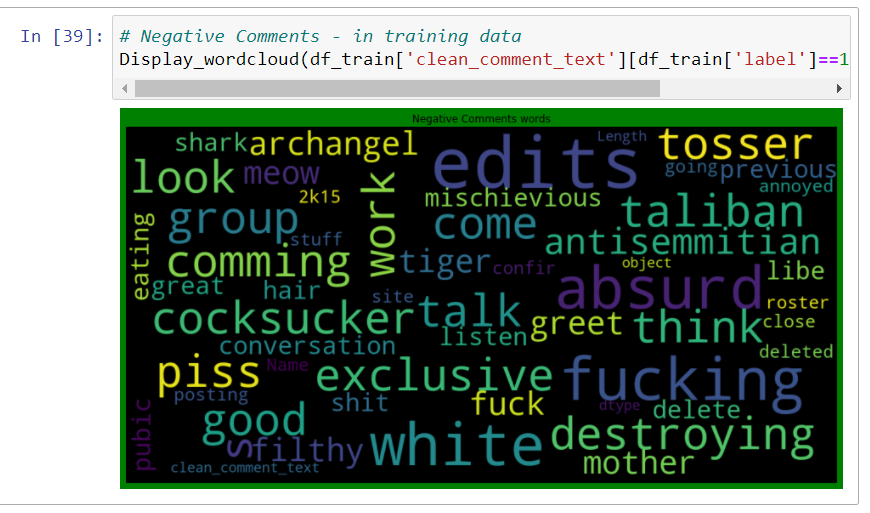
EDA was performed by creating valuable insights using various visualization libraries.



**Malignant Words:**

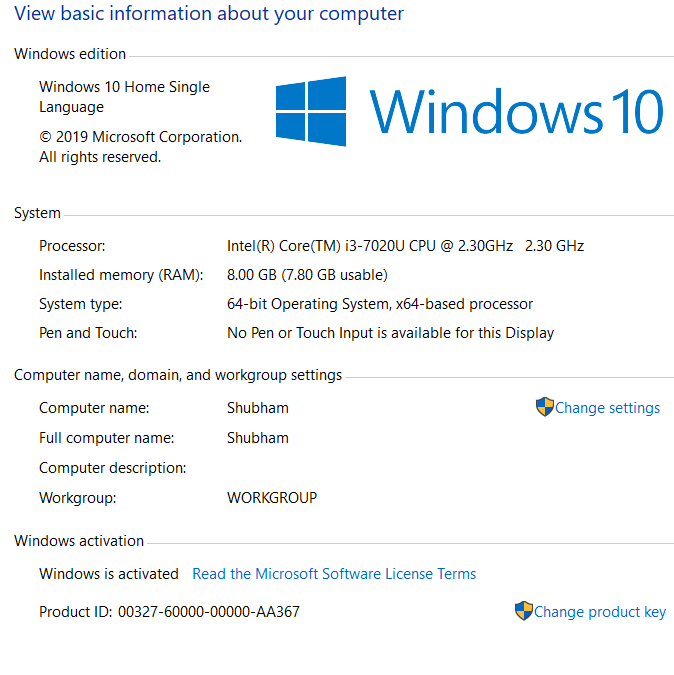


**NoN Malignant Words:**



## HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

***HARDWARE:***



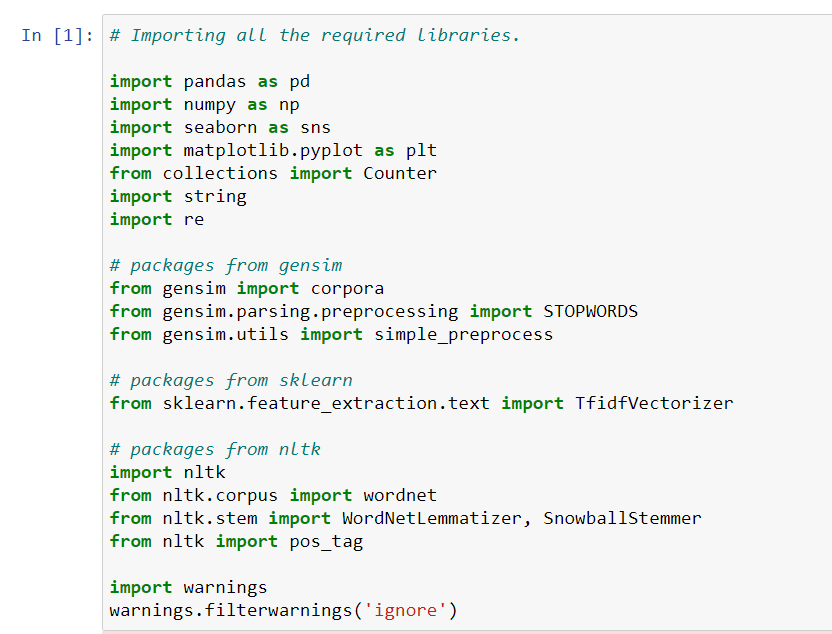
***SOFTWARE:***

Jupyter Notebook (Anaconda 3) – Python 3.7.6

Microsoft Excel 2010

***LIBRARIES:***

The tools, libraries and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy stats, etc.

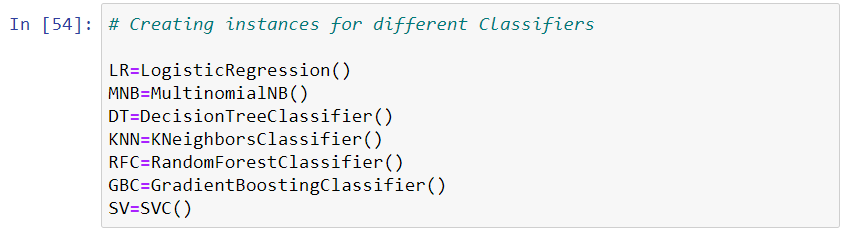


# MODEL/S DEVELOPMENT AND EVALUATION

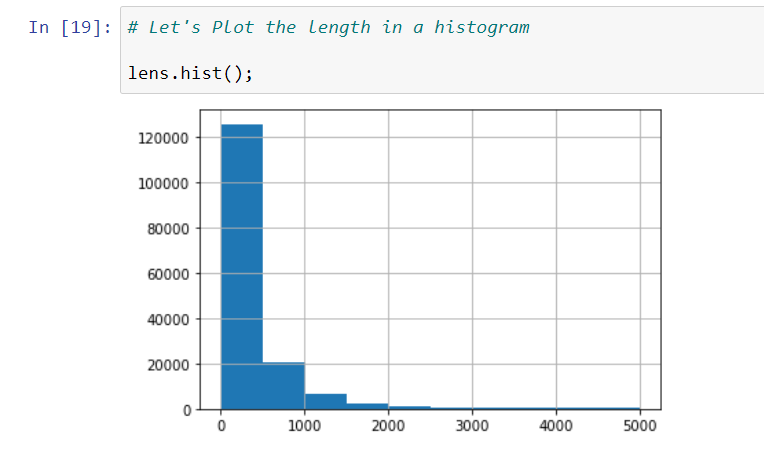
## IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS)

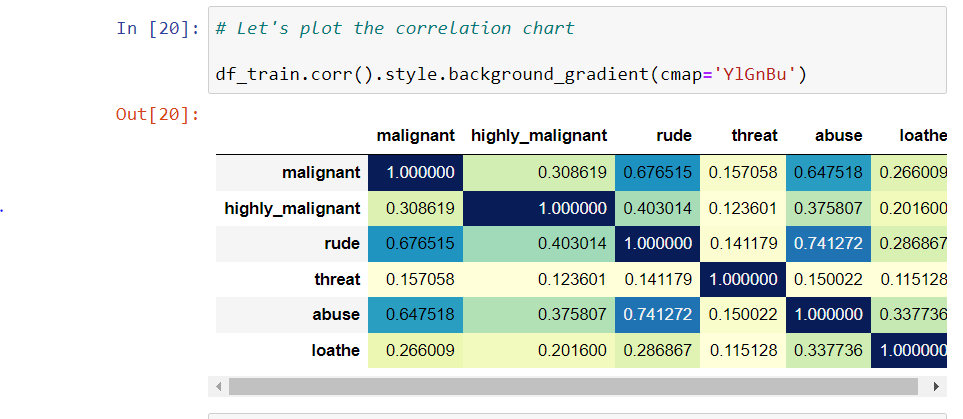
The dataset is loaded and stored in a data frame. We need to perform some text processing to remove unwanted words and characters from our text. I used the nltk library and the string library. Then the data was analysed and visualized to extract insights about the comments. The sentence in the cleaned data, were broken down into vectors using Tokenizer from Keras and each word was converted into sequence of integers. Comments are variable in length, some are one-word replies while others are vastly elaborated thoughts. To overcome this issue, we use Padding. With the help of padding, we can make the shorter sentences as long as the others by filling the shortfall by zeros, and on the other hand, we can trim the longer ones to the same length as the short ones [3]. I used the “pad\_sequences” function from the “Keras” library and, I fixed the sentence length at 200 words and applied pre padding (i.e. for shorter sentences, 0’s will be added at the beginning of the sequence vector) A model was built using Keras and Tensorflow. For our classification task, I used both CNN and LSTM neural networks. The model consisted of Embedding layer, which is responsible for embedding. MaxPool layer used to focus on the important features. Bi-directional LSTM was used for one forward and one backward network. Last layer consisted of Sigmoid layer, which will predict probabilities for each kind of features in our dataset. The training dataset was split into training and validation set. 20% of the training data was kept aside for validation. The model was compiled with various optimizers, amongst which adam performed better and metrics like loss and AUC were used to evaluate the model. The dataset was then fit on training data and validated on validation dataset. It gave a quite good AUC of about 98.3% with 2 epochs. The loss was also decreasing significantly with increase in epoch, and finally the model was used to predict on the testing dataset.

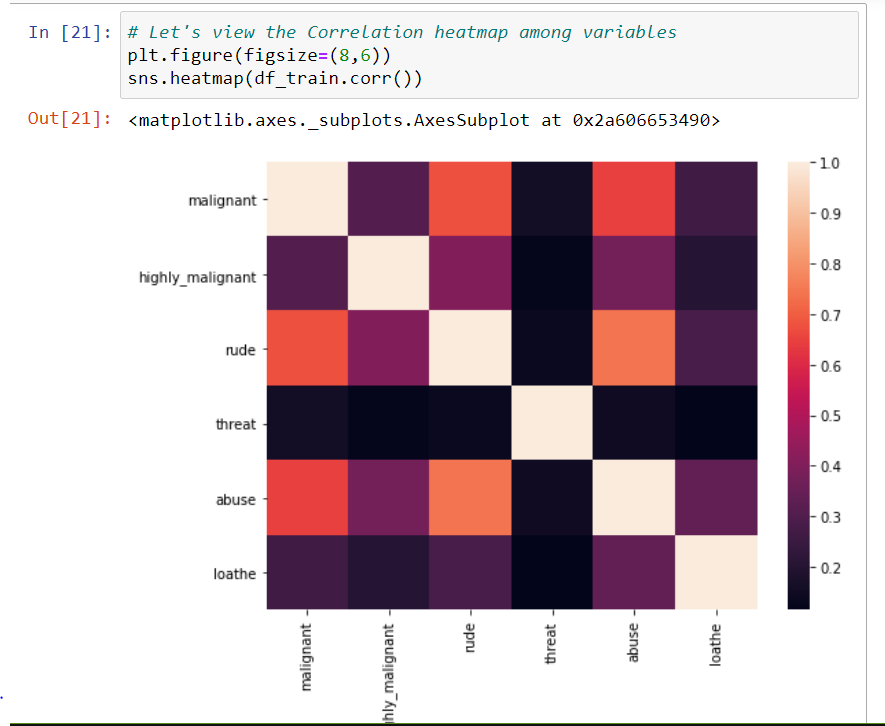
## TESTING OF IDENTIFIED APPROACHES (ALGORITHMS)

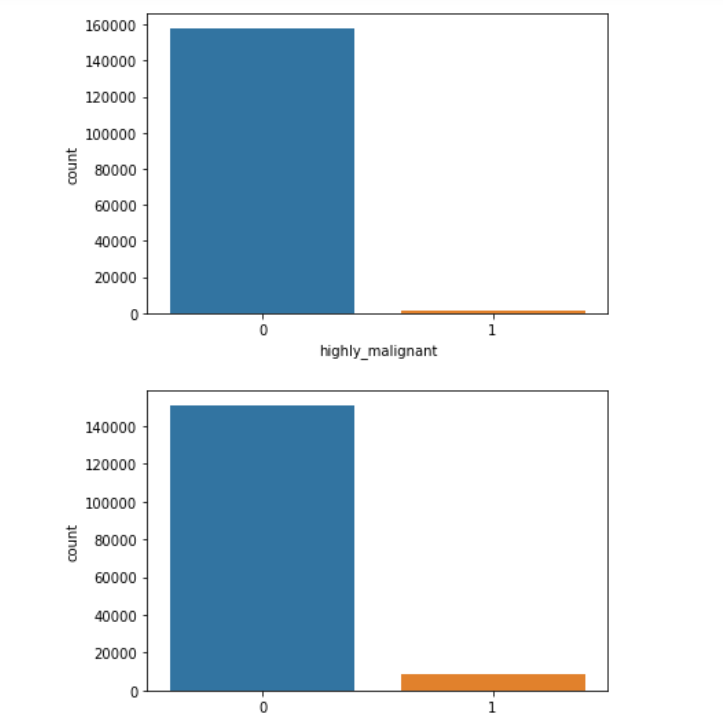
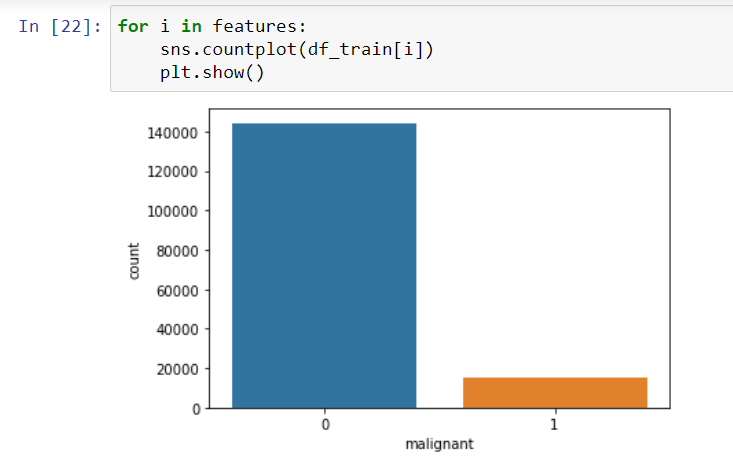


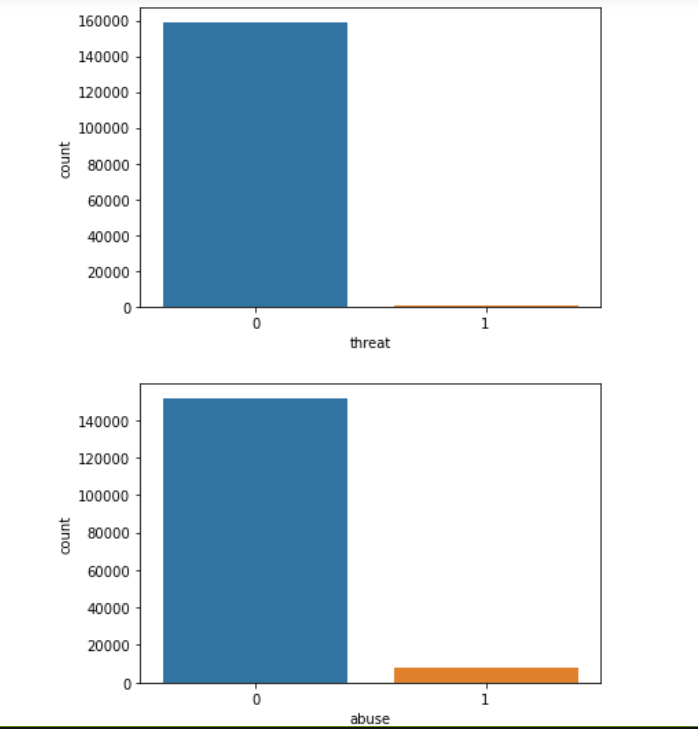
## VISUALIZATIONS

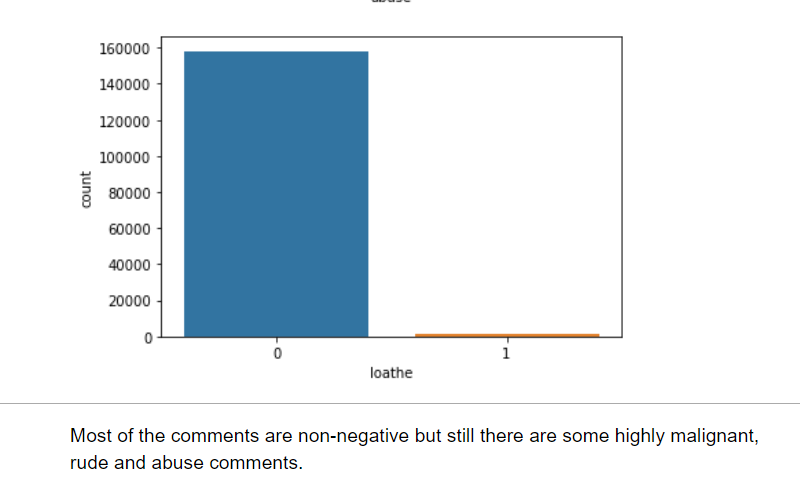




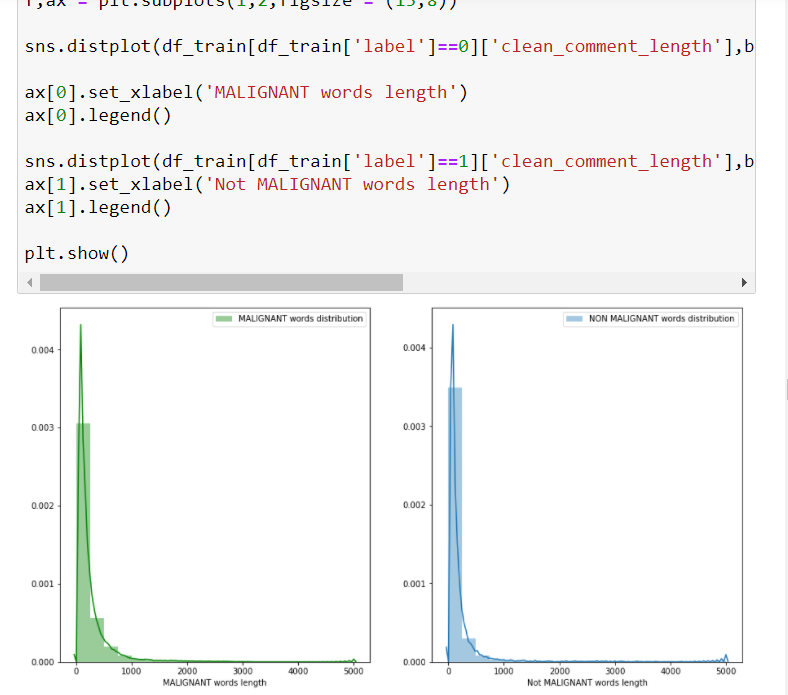




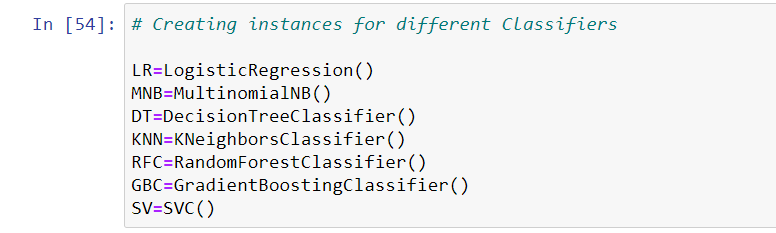


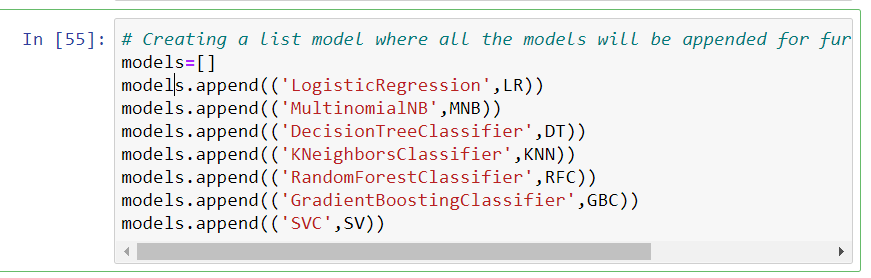


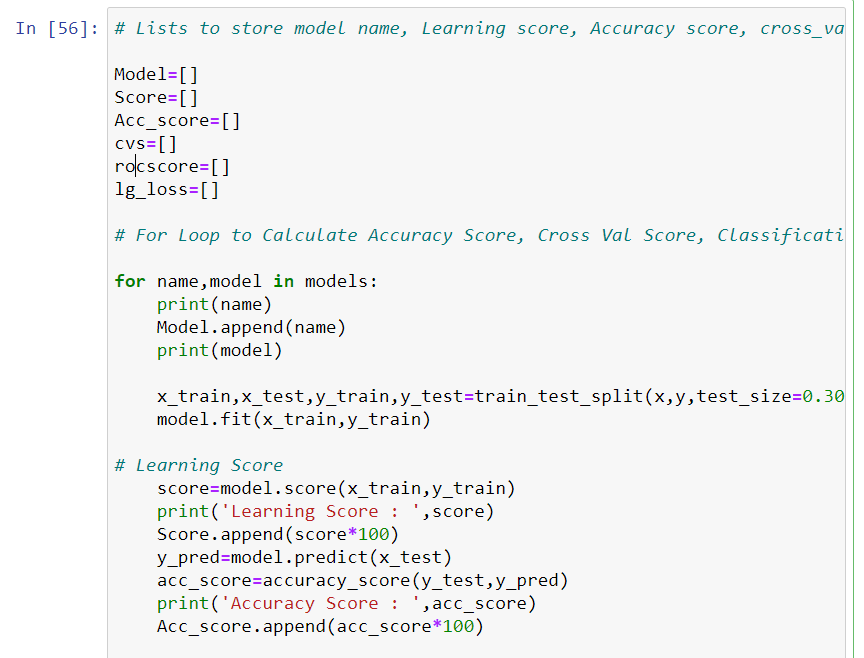


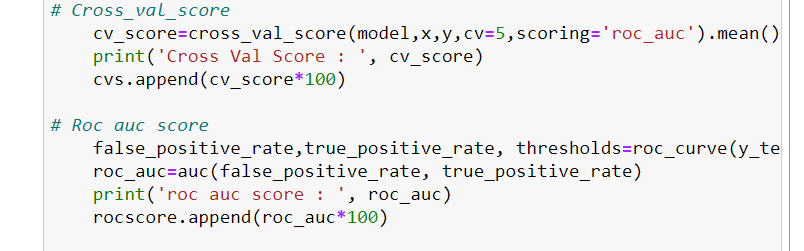


***RUN AND EVALUATED SELECTED MODELS***

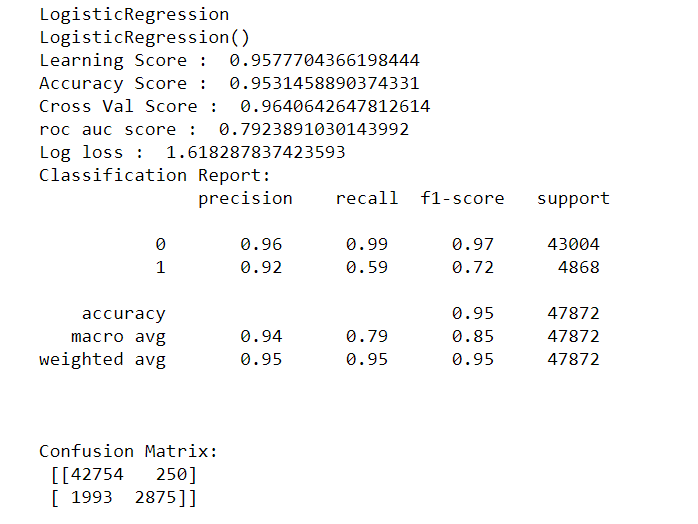


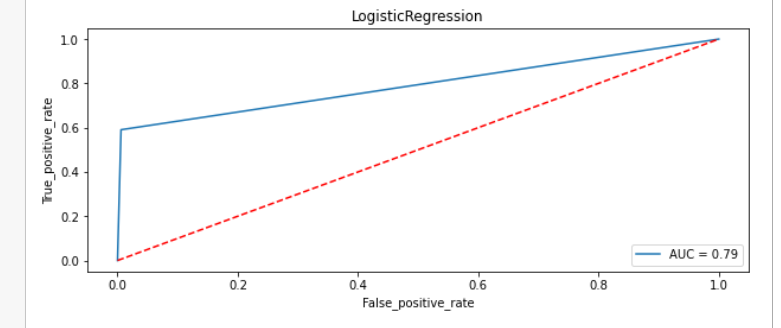


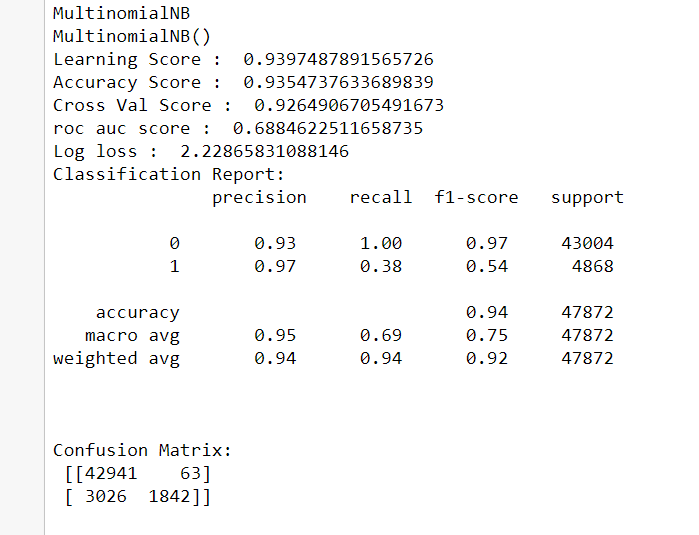


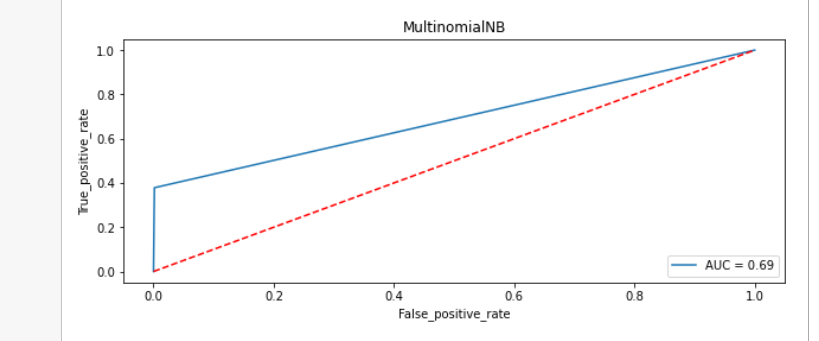


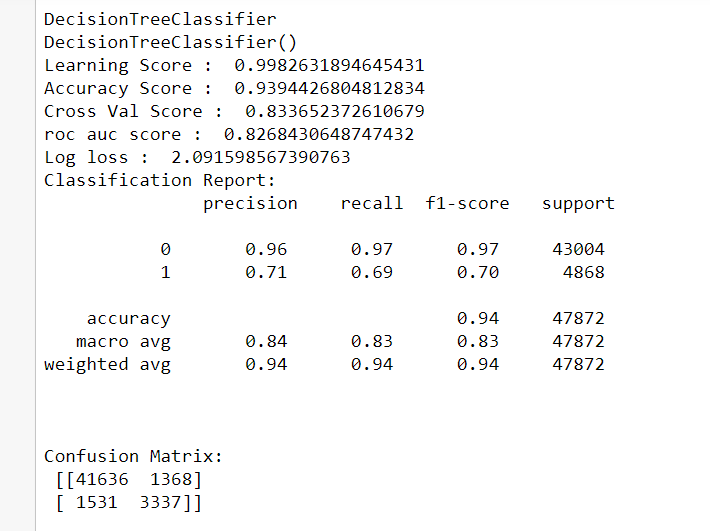


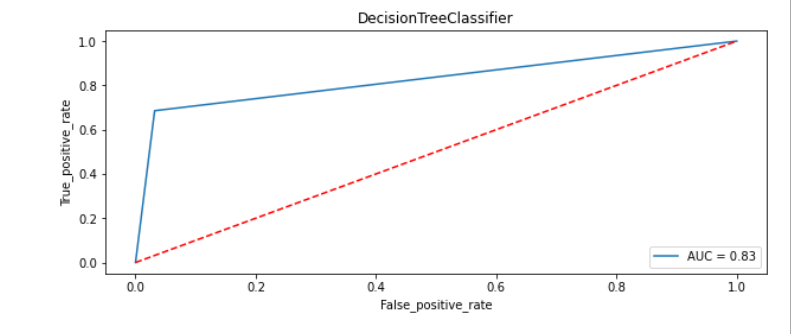


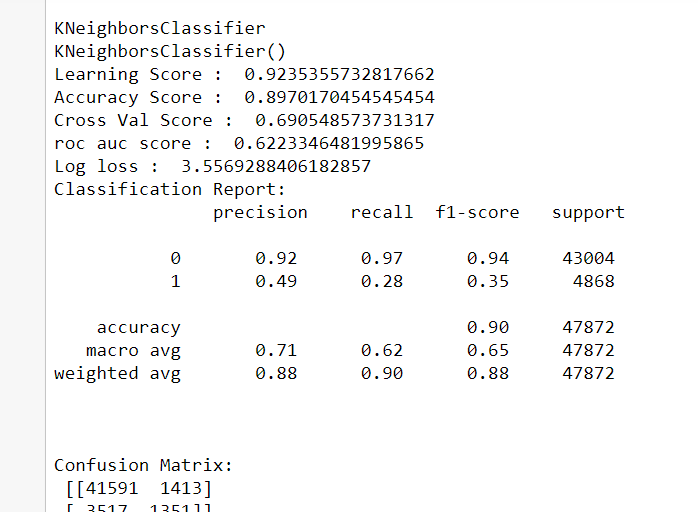


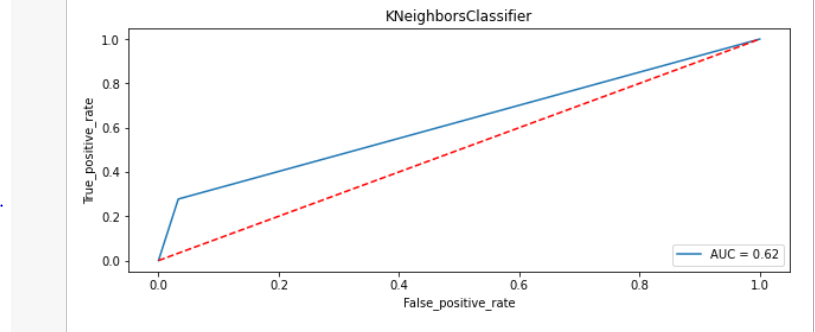


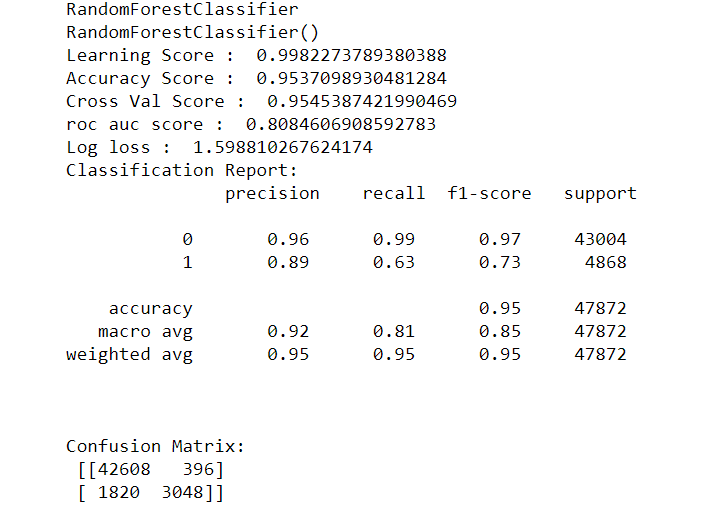


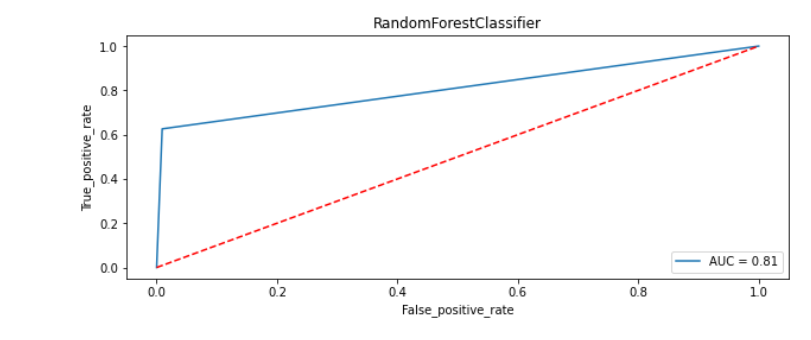


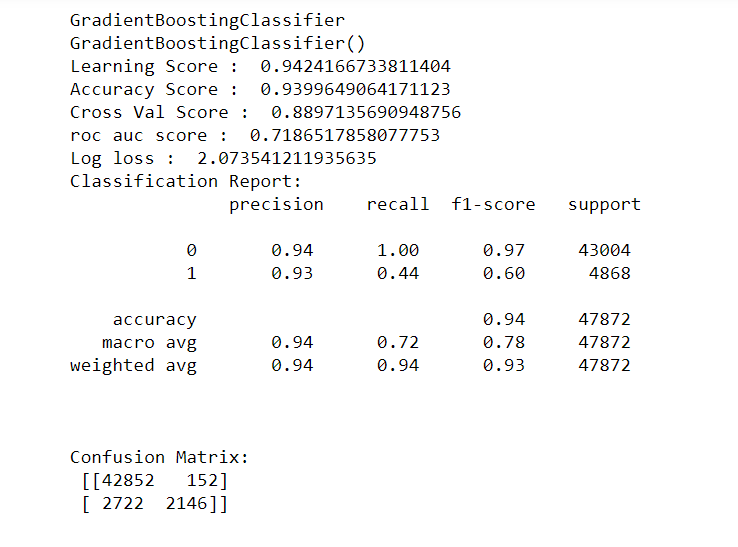


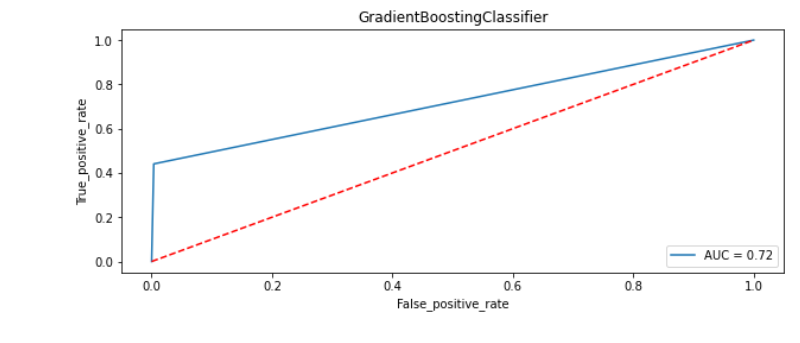


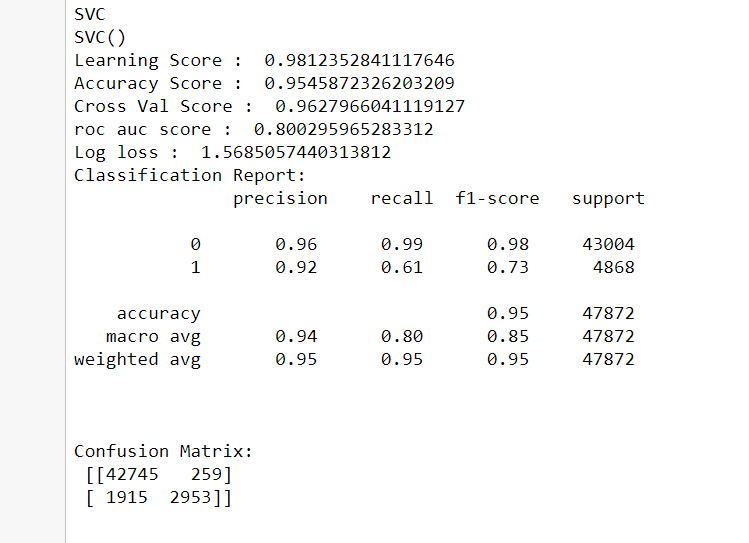


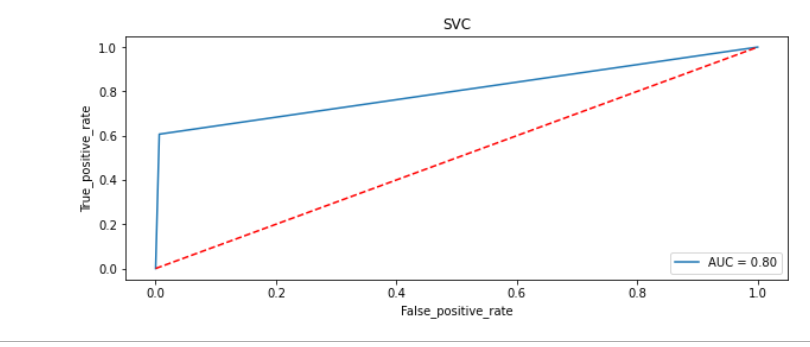


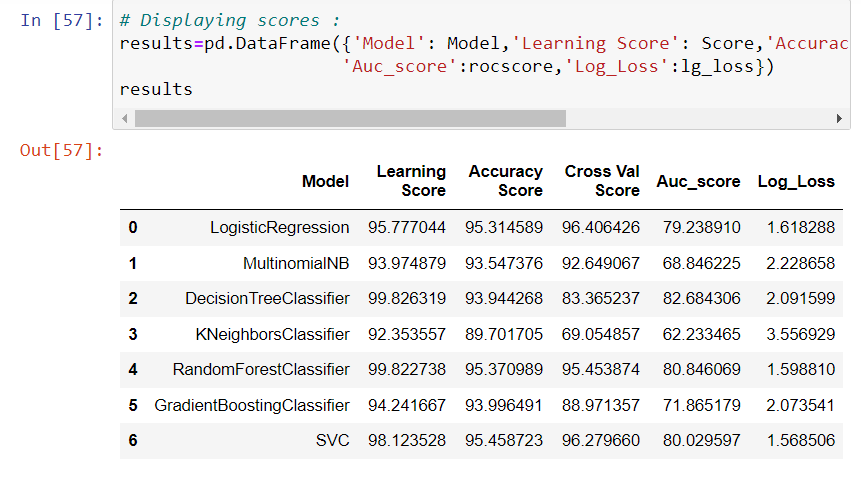


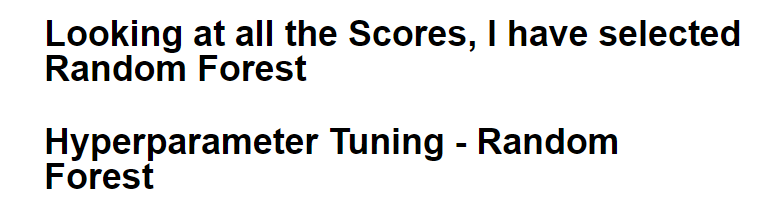






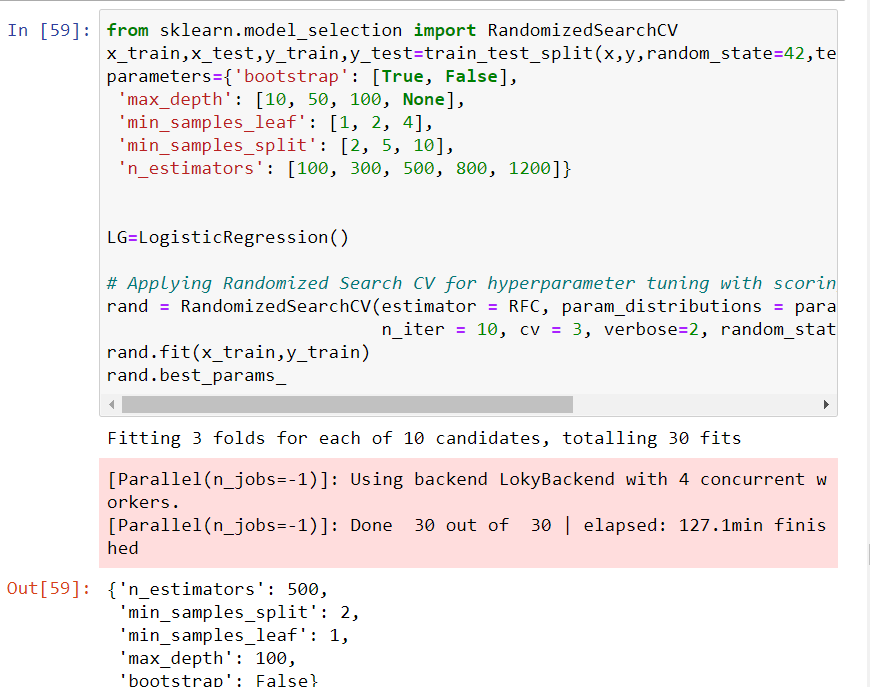


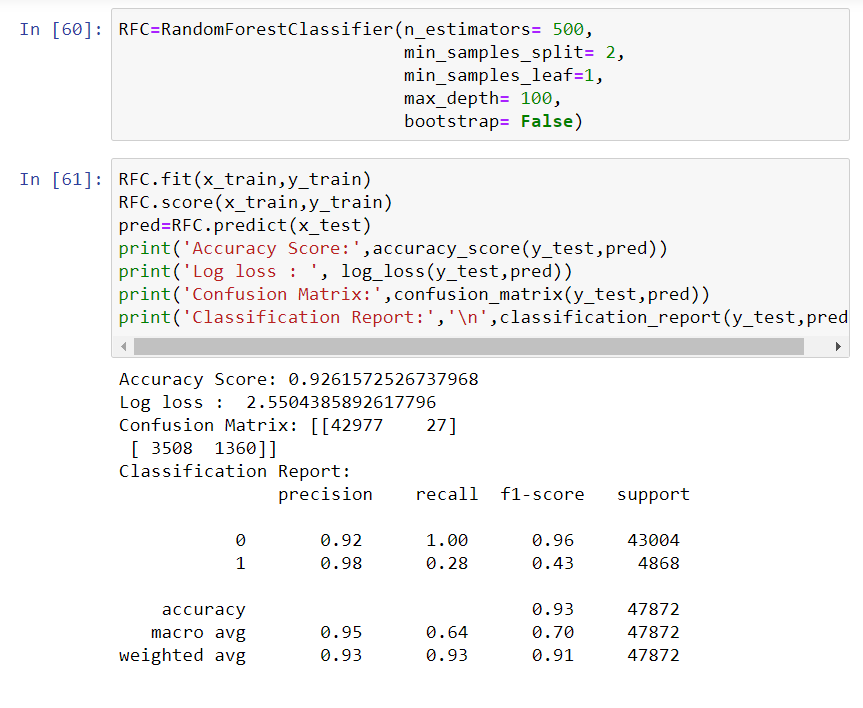


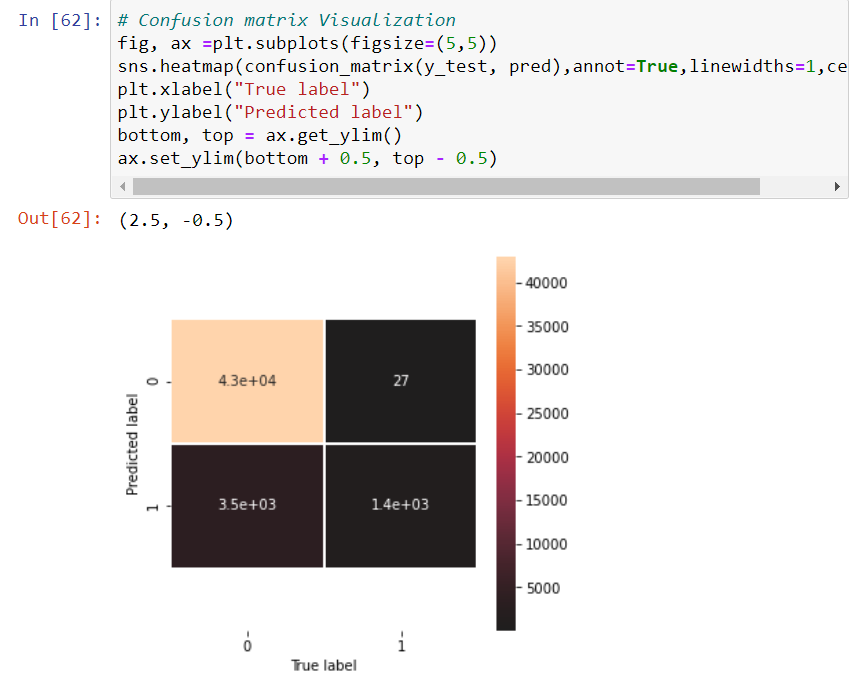


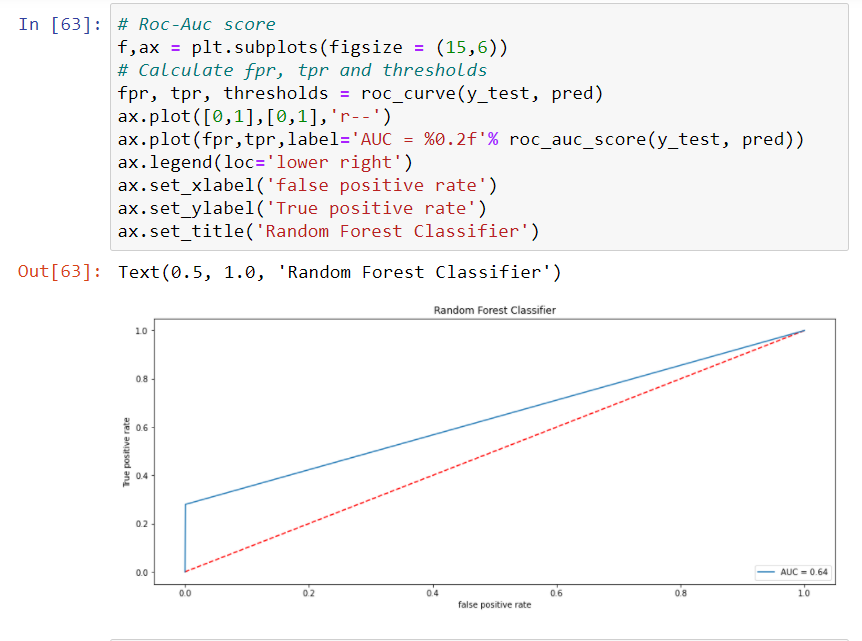
**INTERPRETATION OF THE RESULTS**

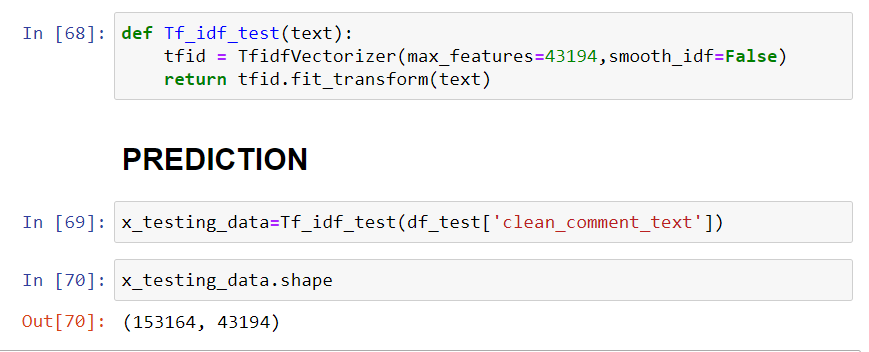
**FINAL MODEL**

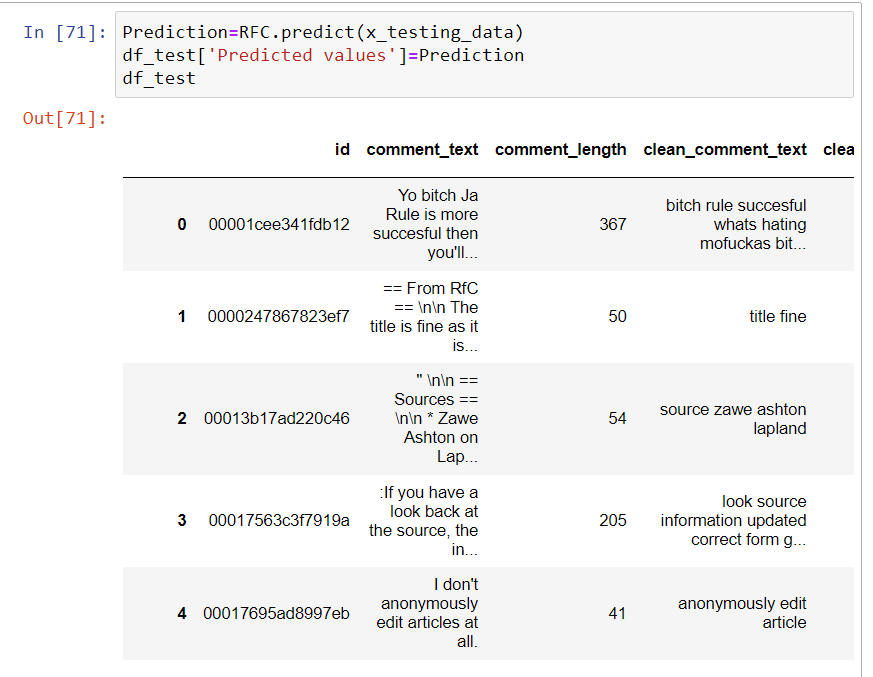




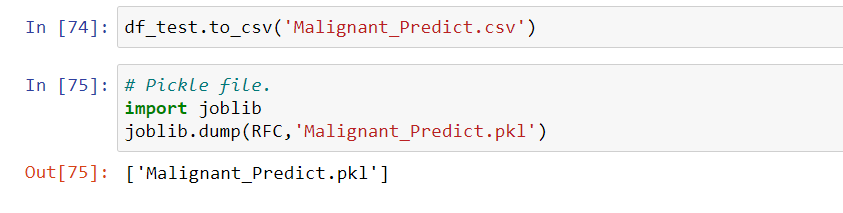












# CONCLUSION

## 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.