

**MALIGNANT COMMENTS CLASSIFIER PROJECT**

**Submitted by: Sudhanshu Kumar**

# *ACKNOWLEDGMENT*

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot.

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

* [https://medium.com/@dobko\_m/nlp-text-data-cleaning-and-preprocessing- ea3ffe0406c1](https://medium.com/%40dobko_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](https://medium.com/%40cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a) [Chaudhary | Medium](https://medium.com/%40cmukesh8688/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.
* 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:

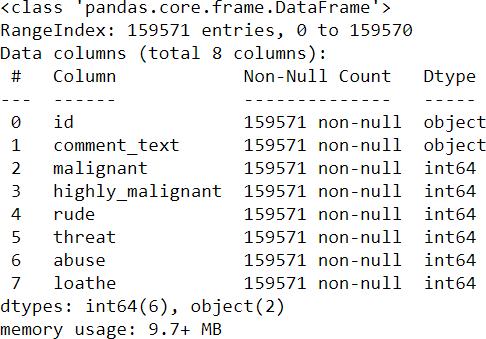
**“** id, comment\_text, “malignant, highly\_malignant, rude, threat, abuse, loathe**”.**

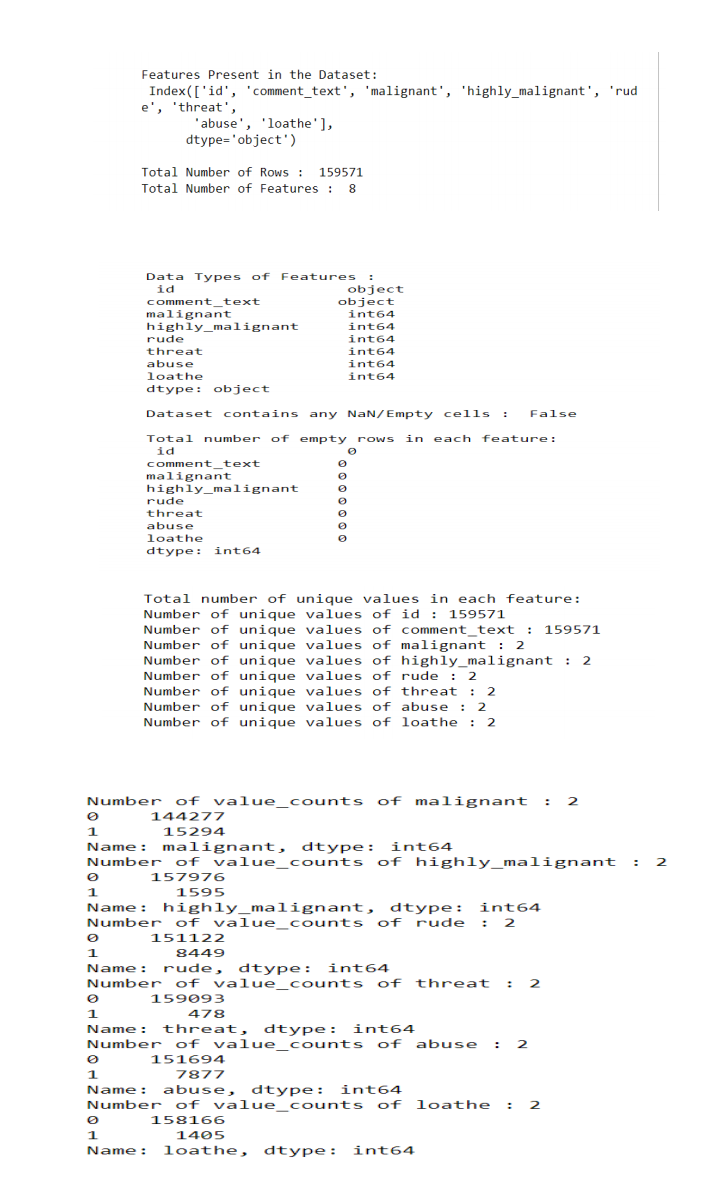
There are 8 columns in the dataset provided:

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.

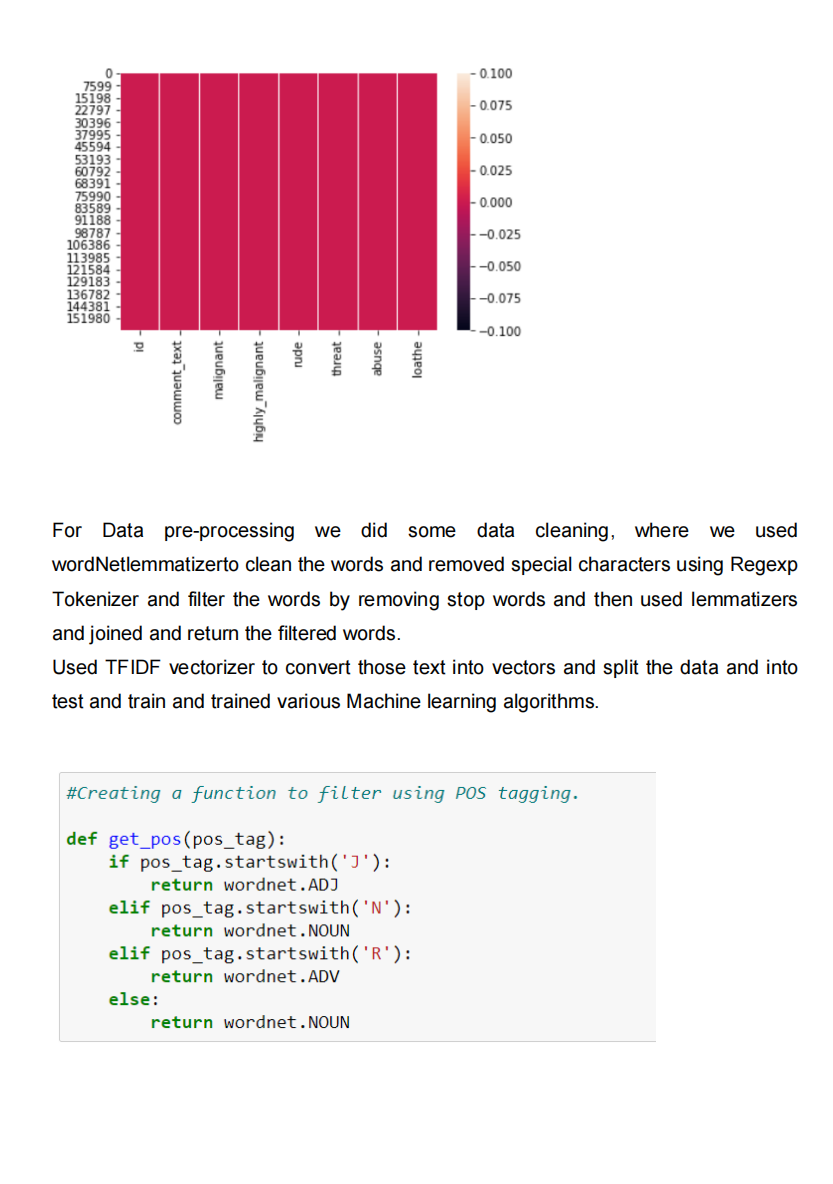
**Comment text:** This column contains the comments extracted from various social media platforms.

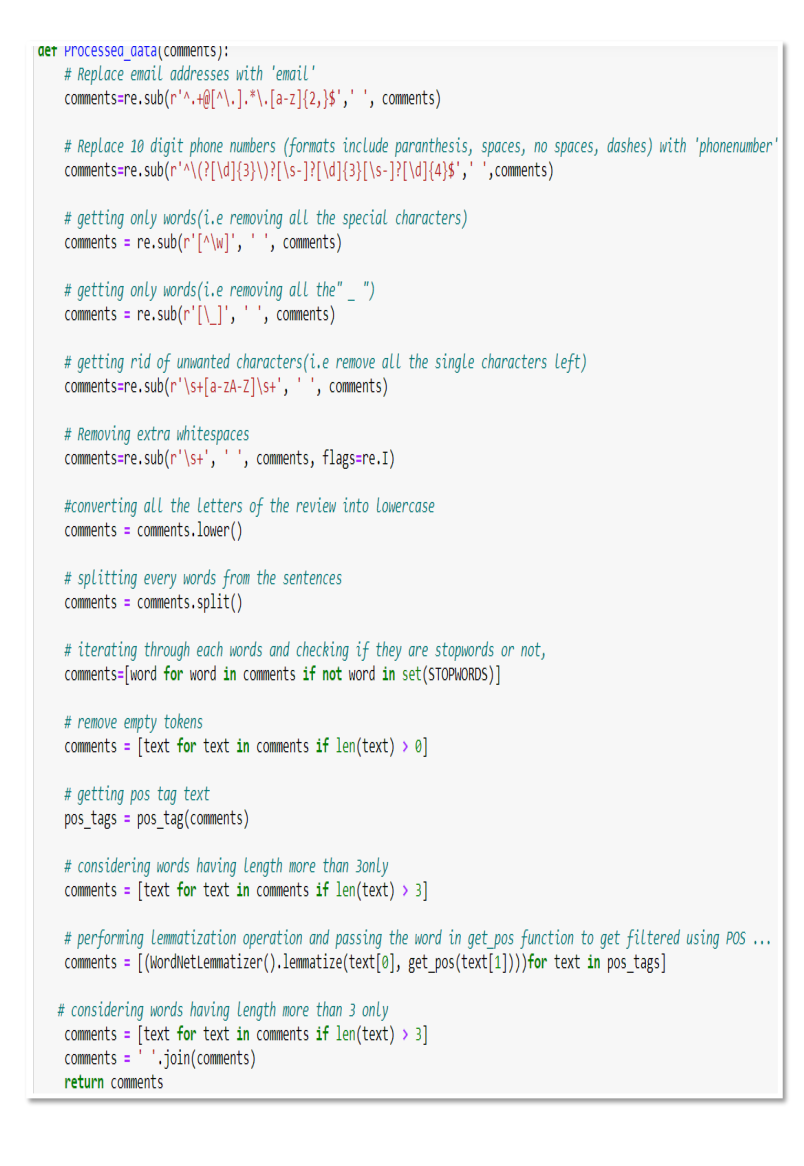


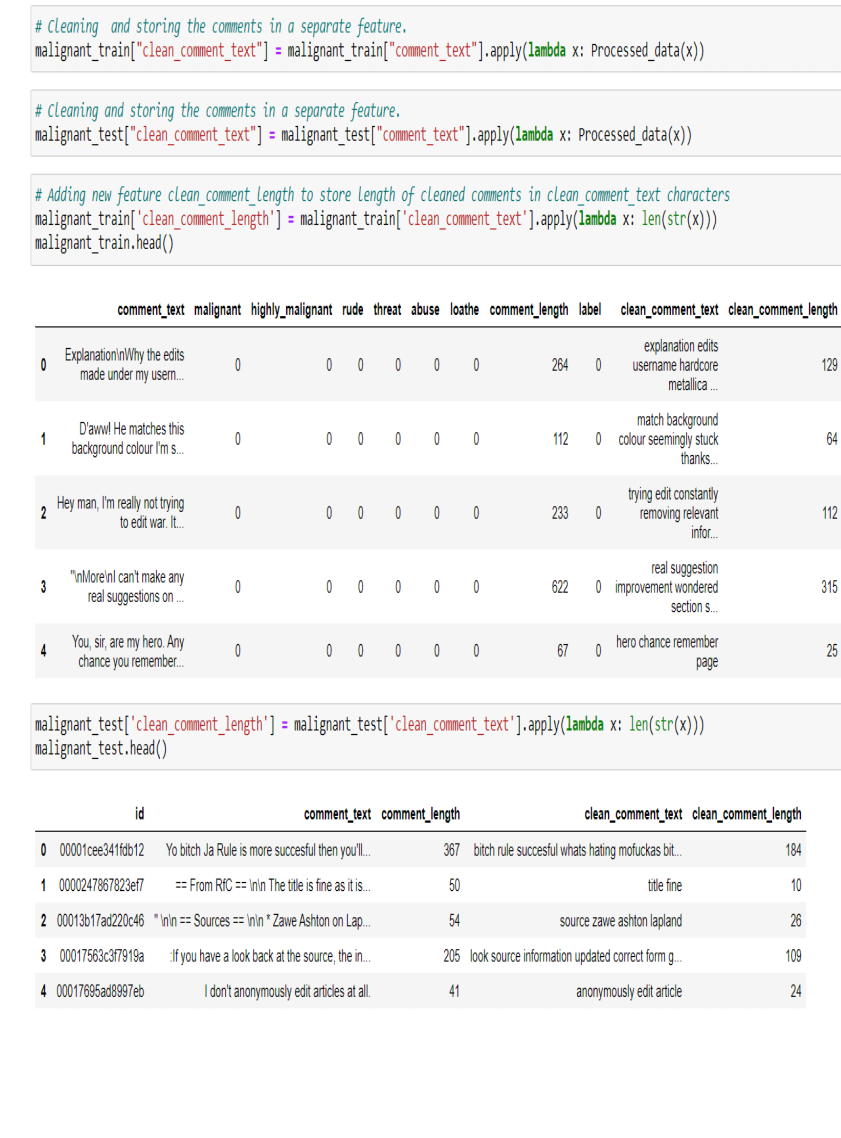


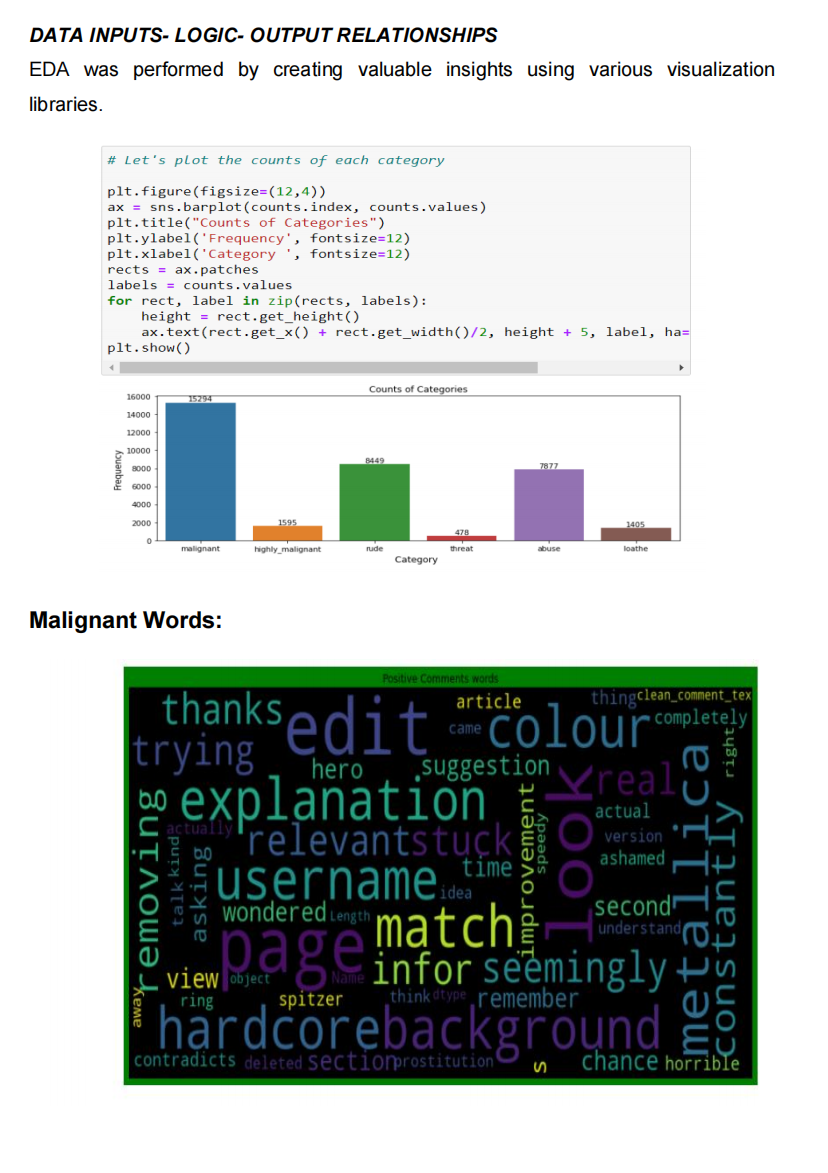


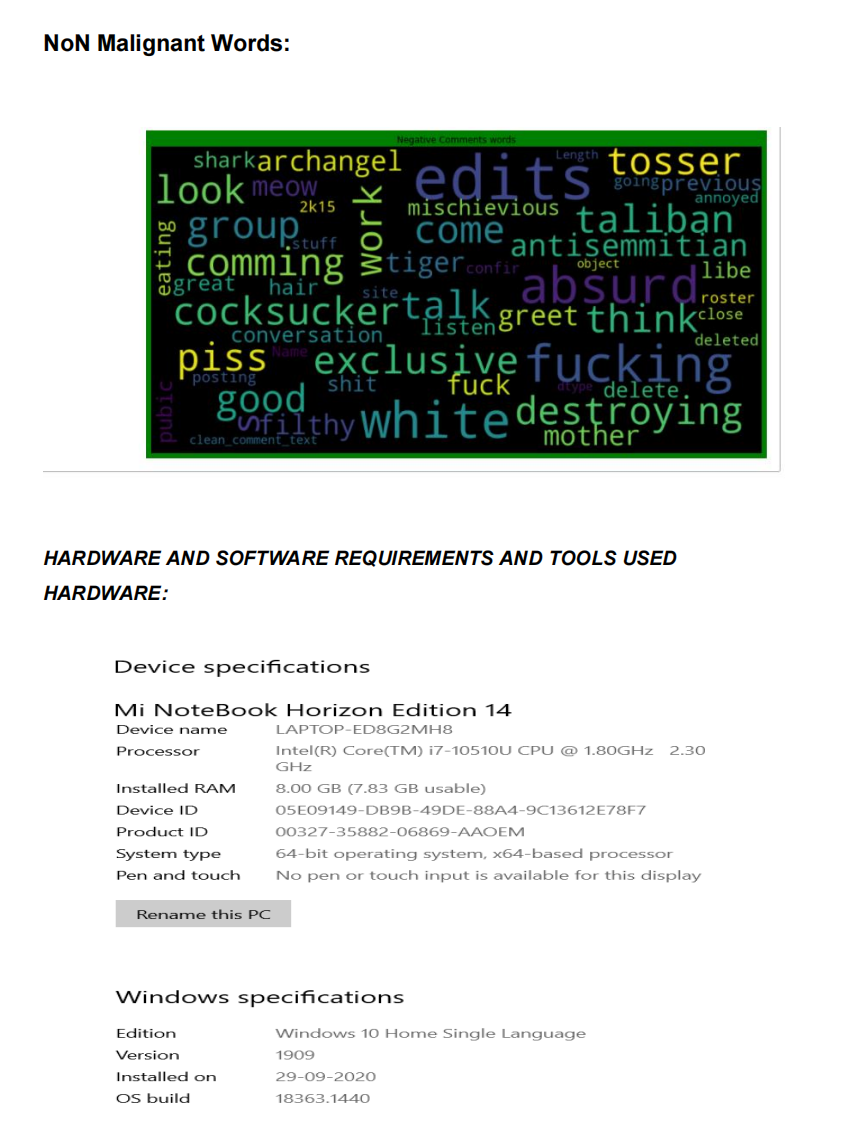


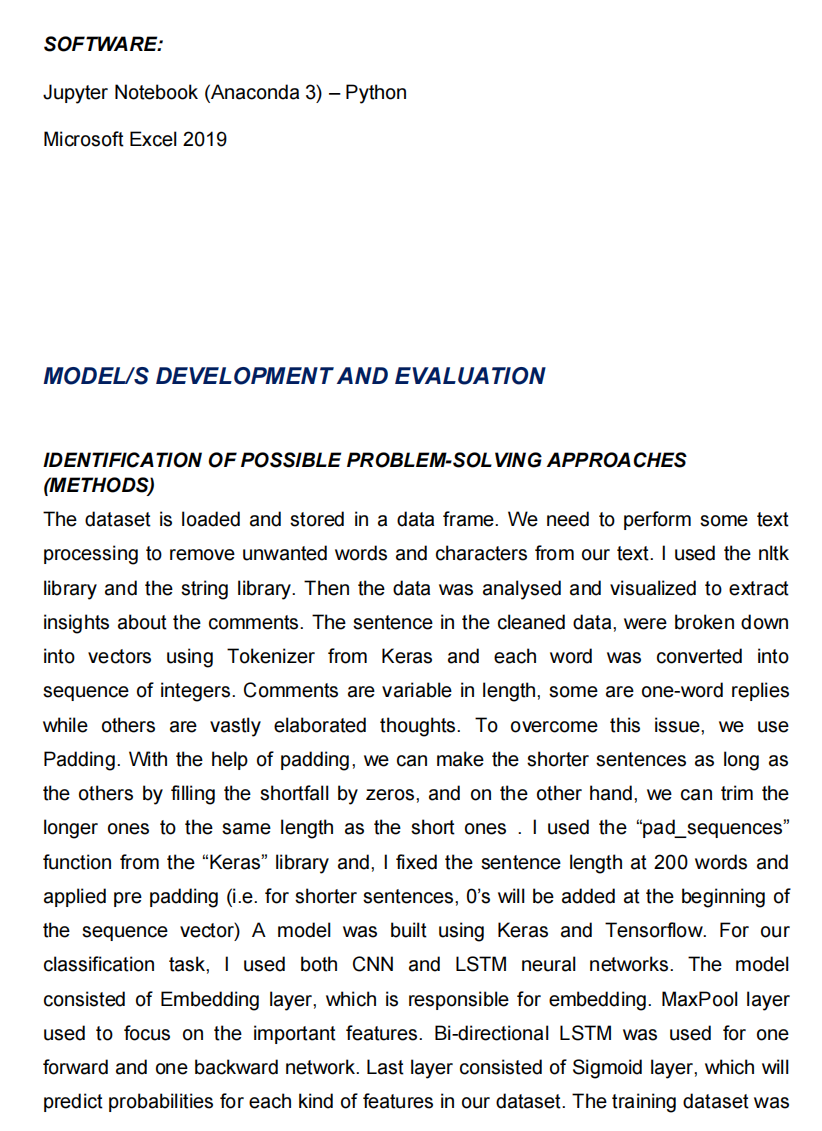


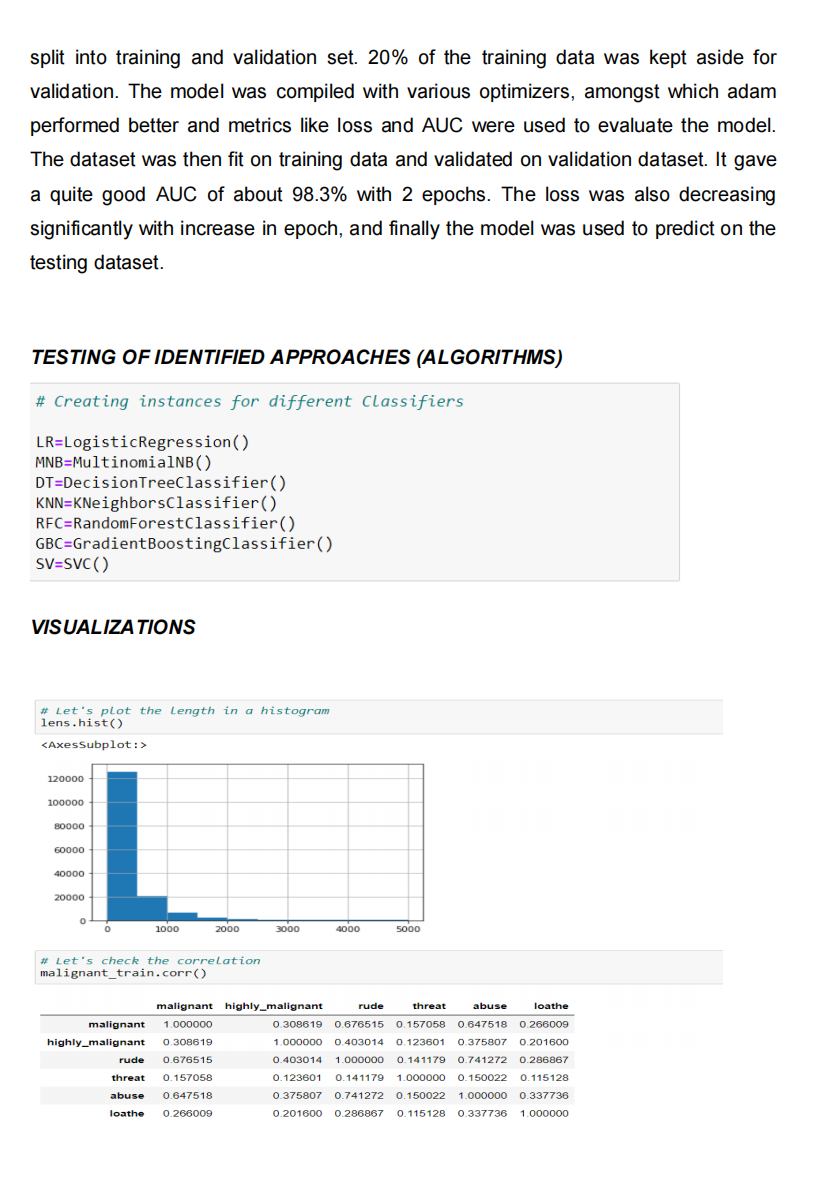


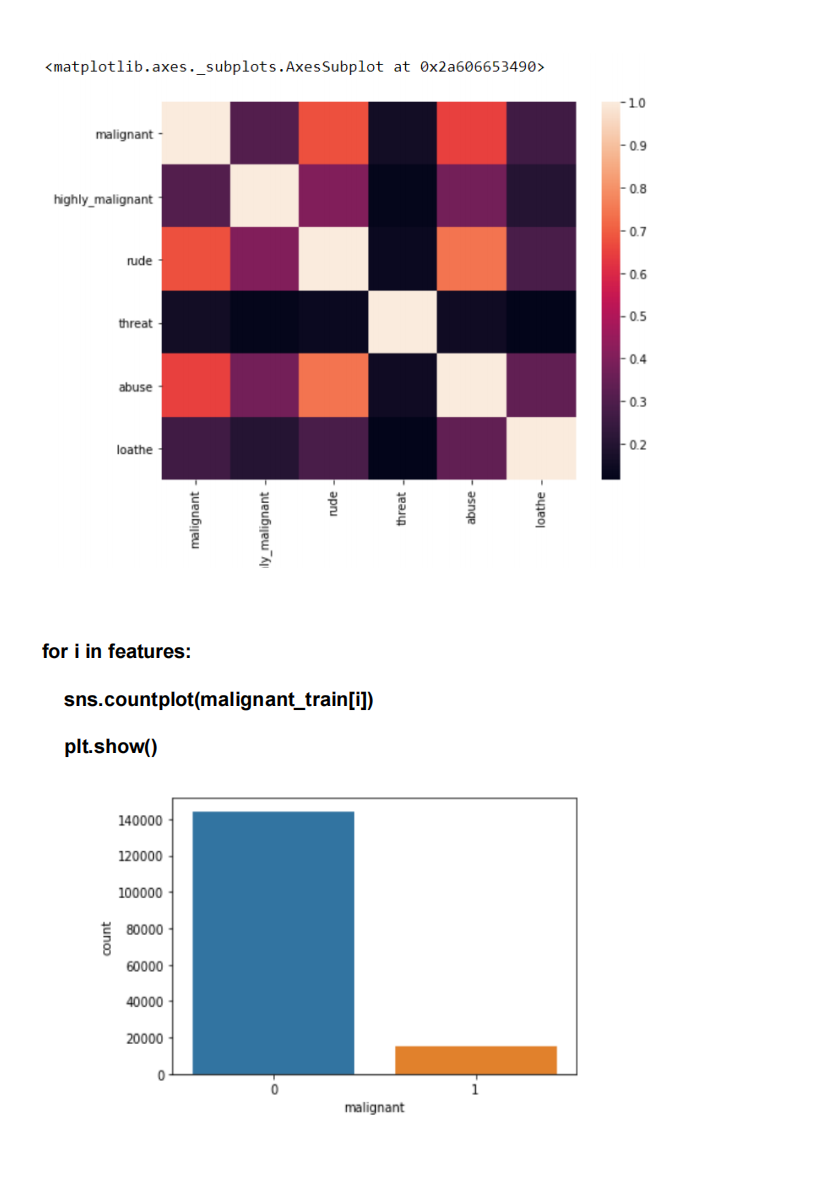


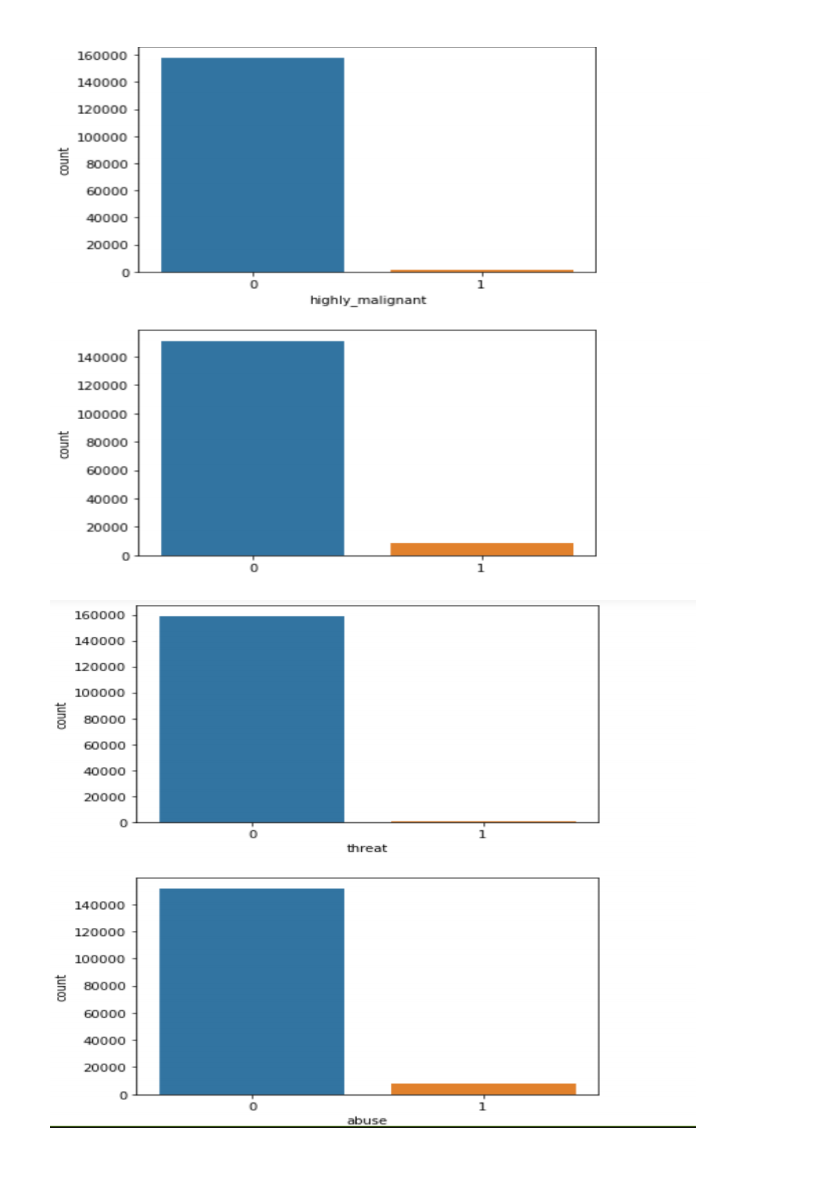


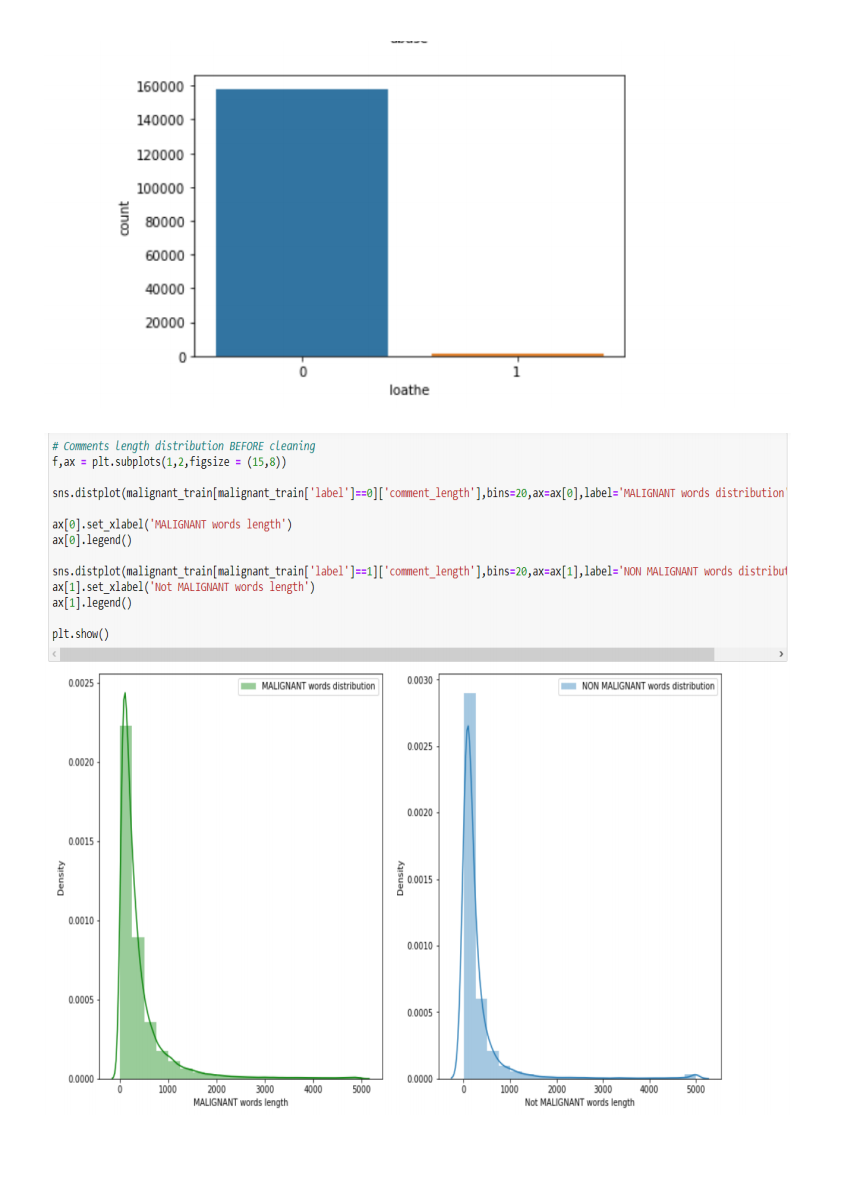


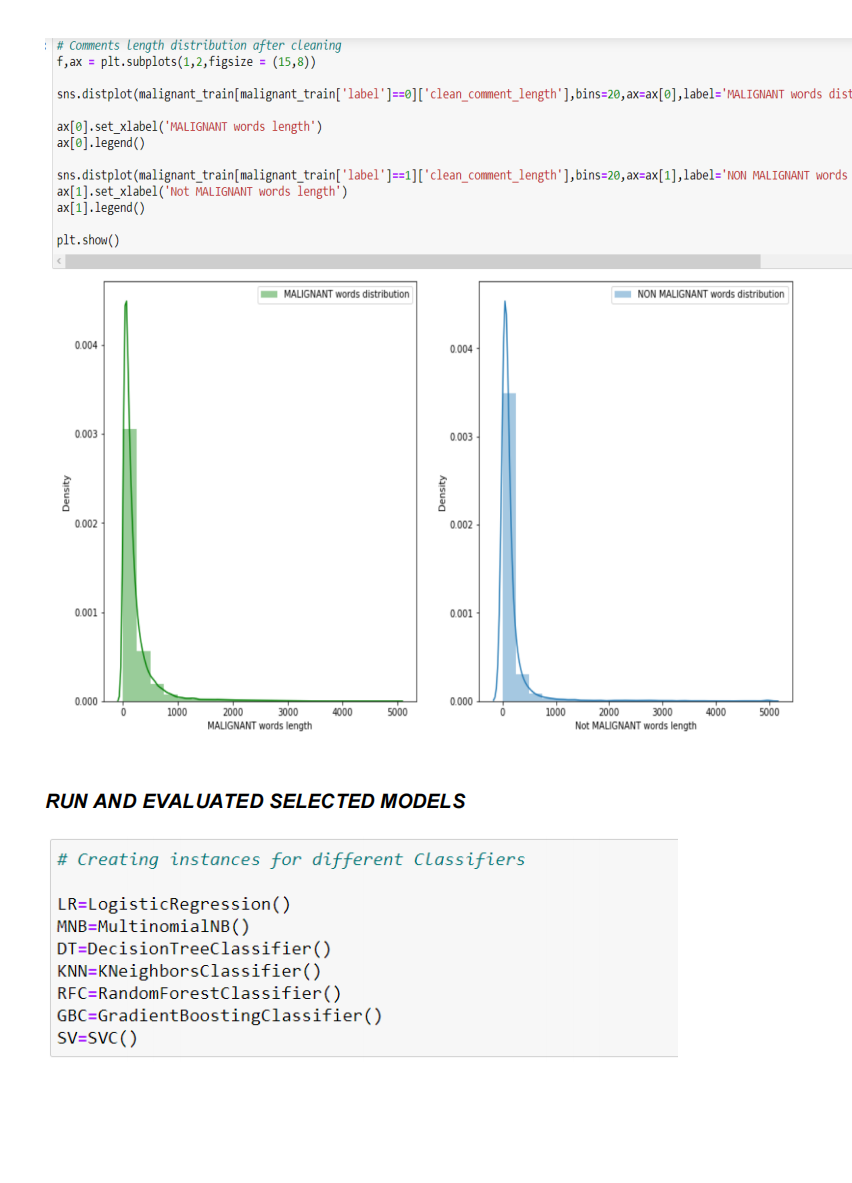




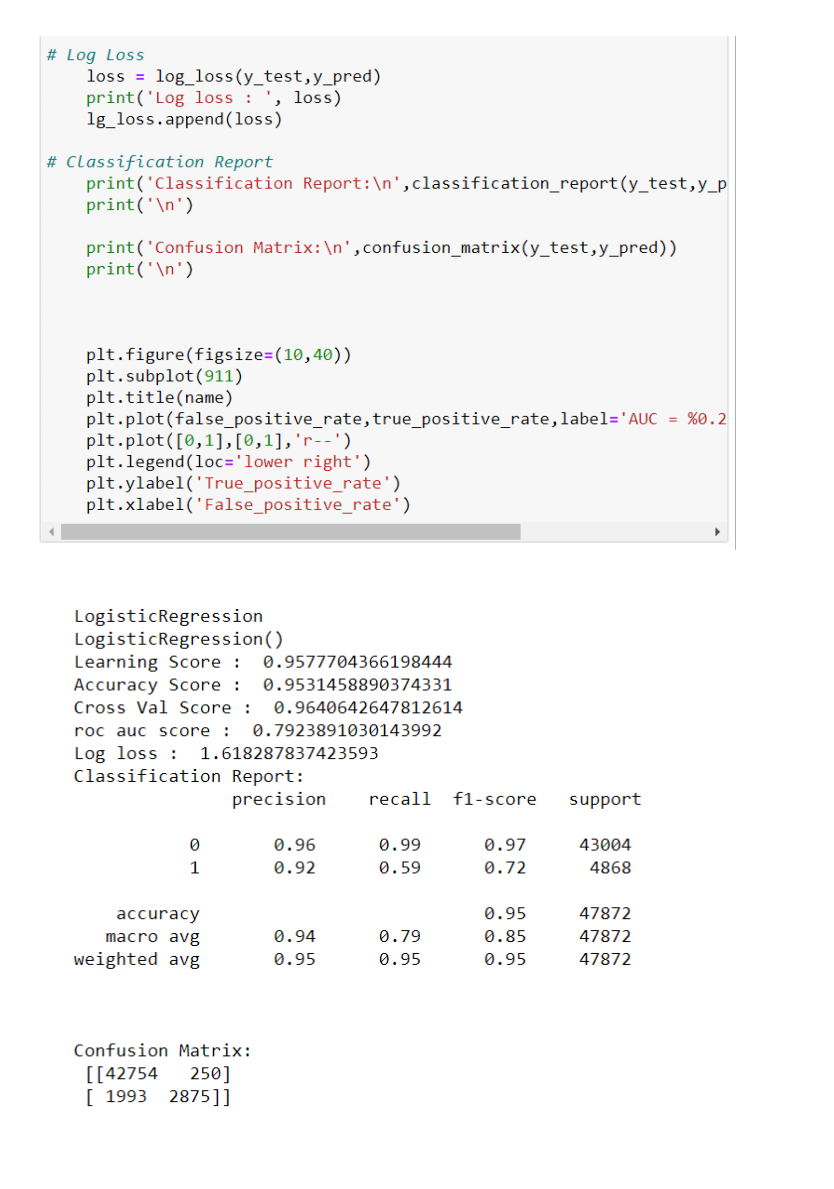


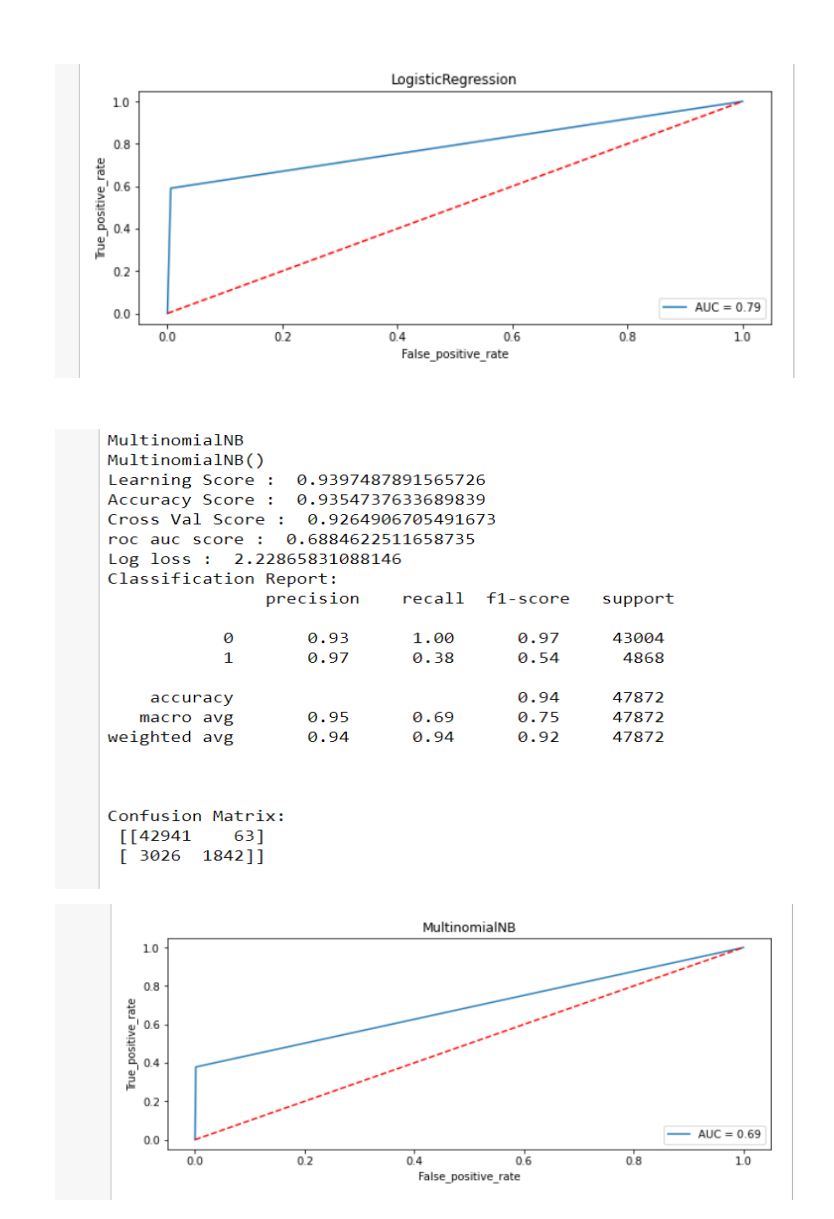




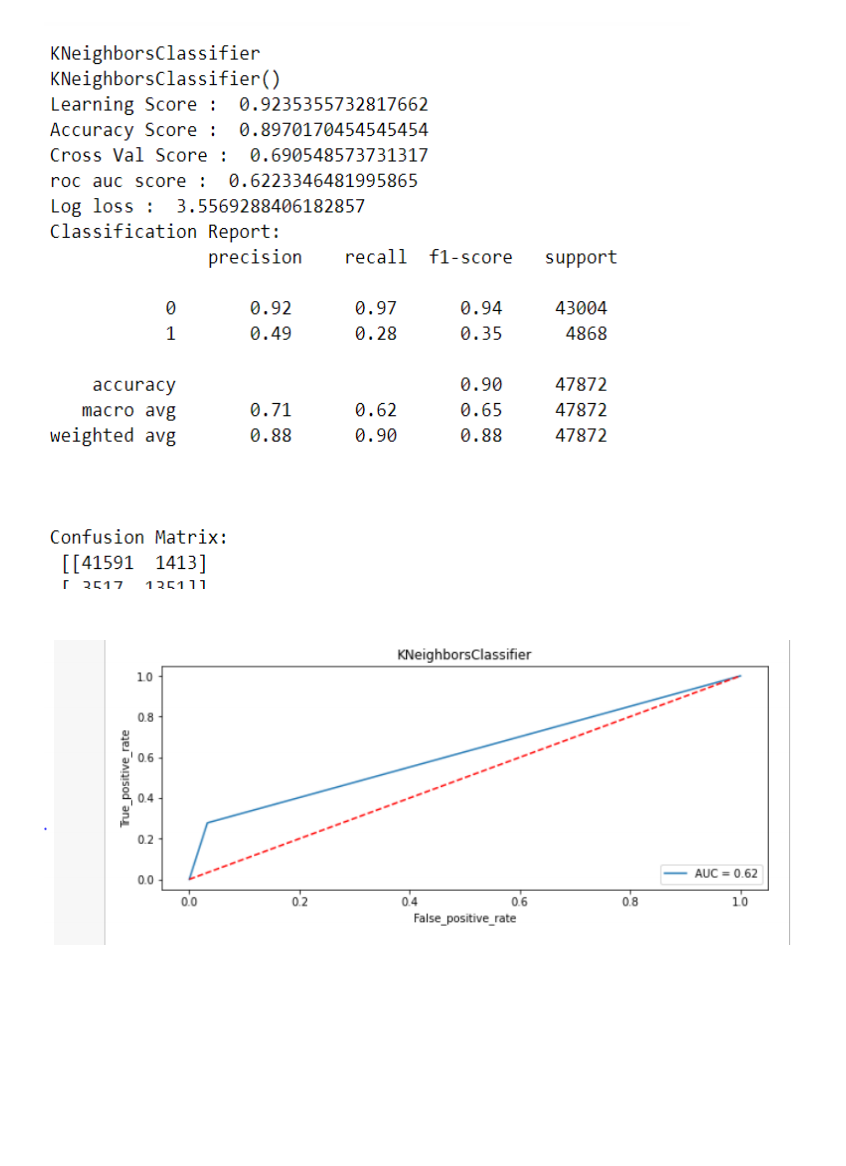


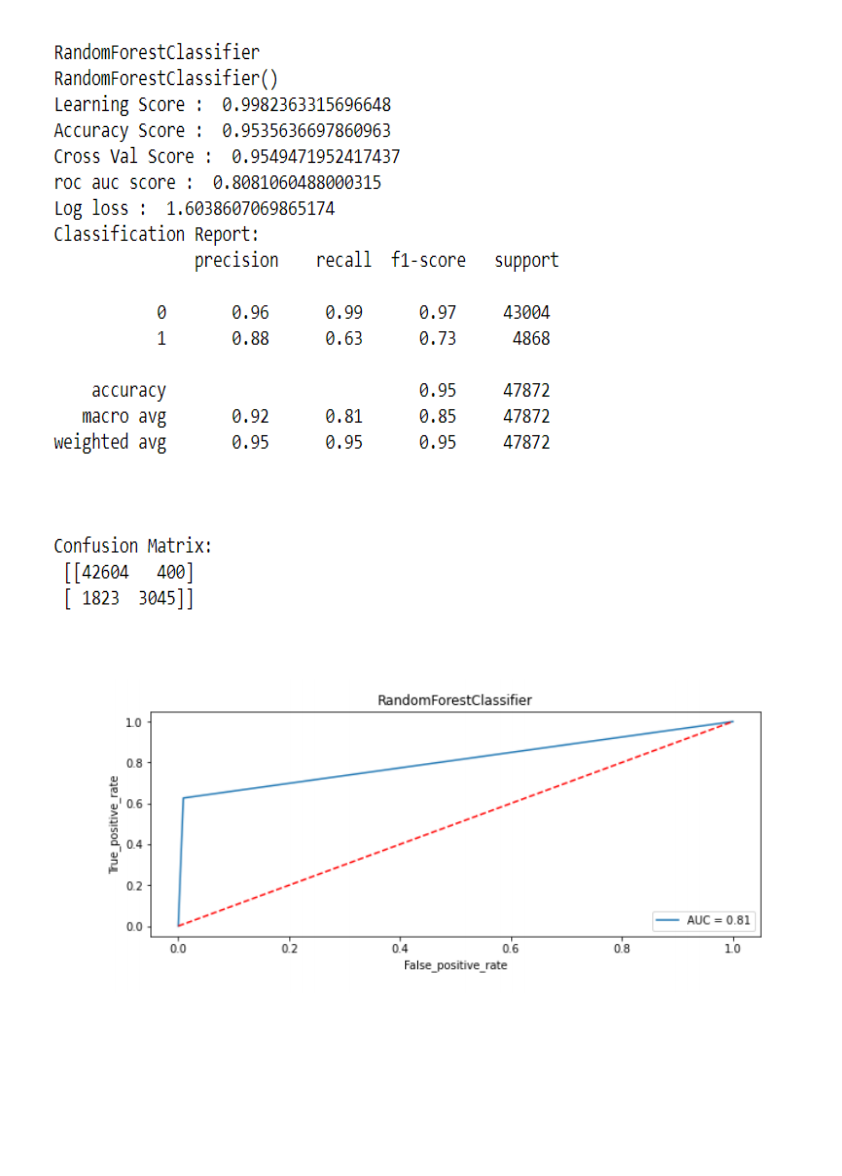


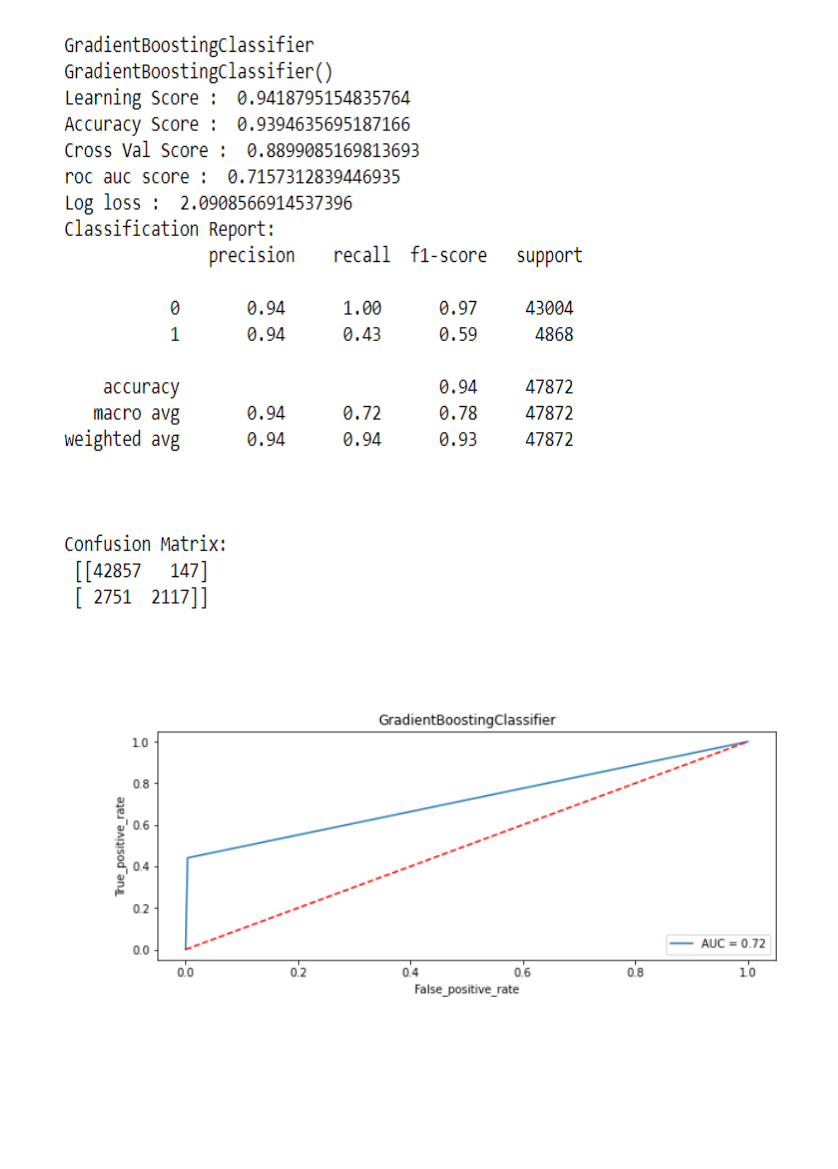


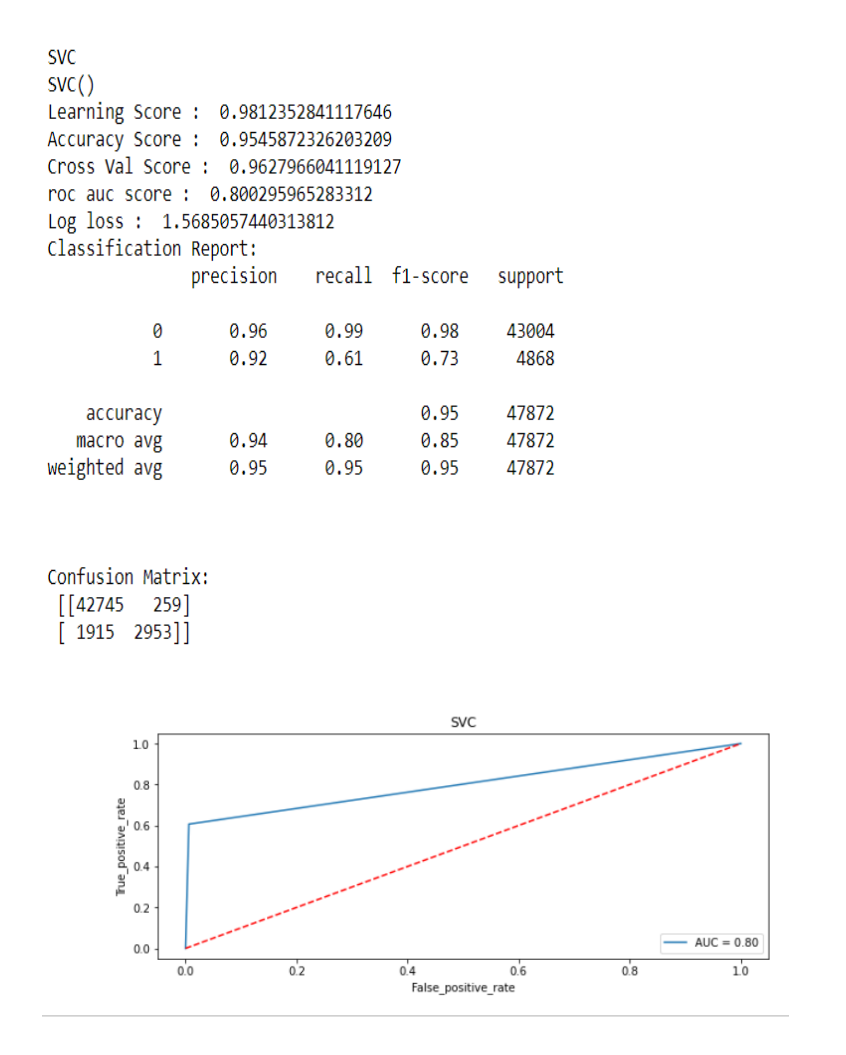


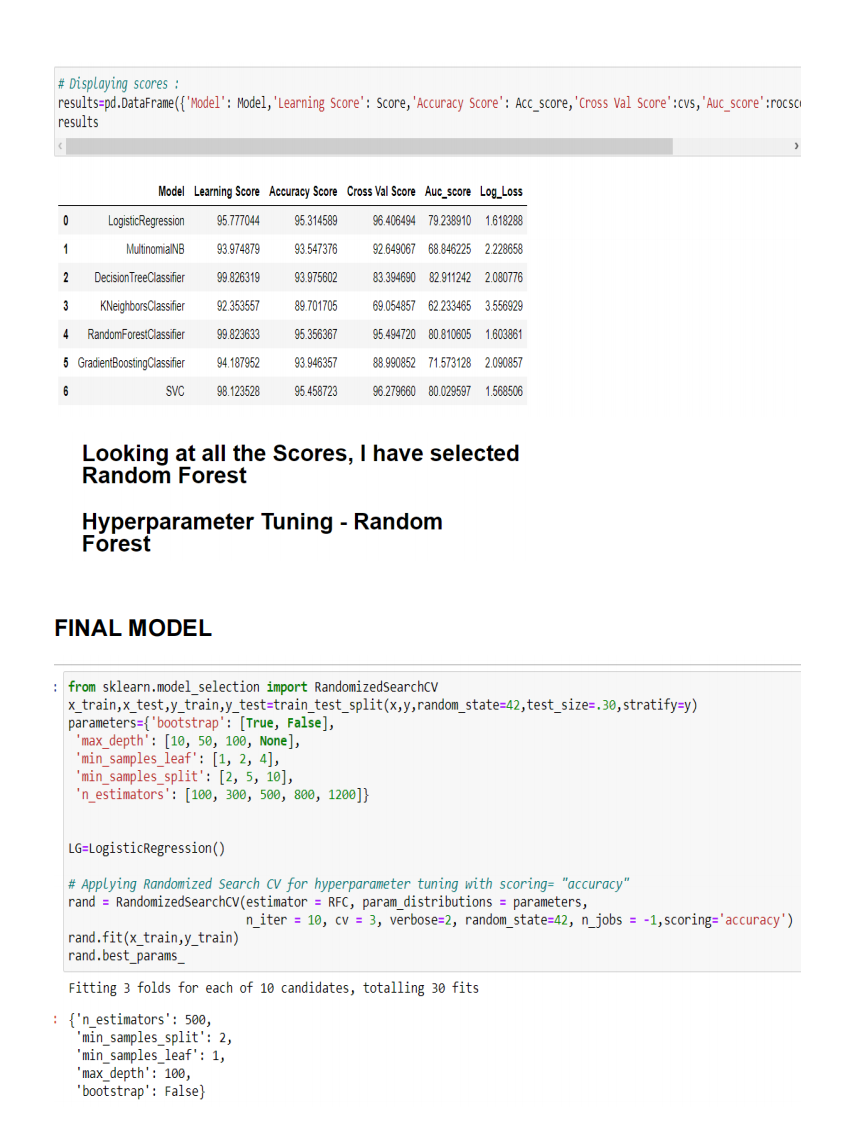




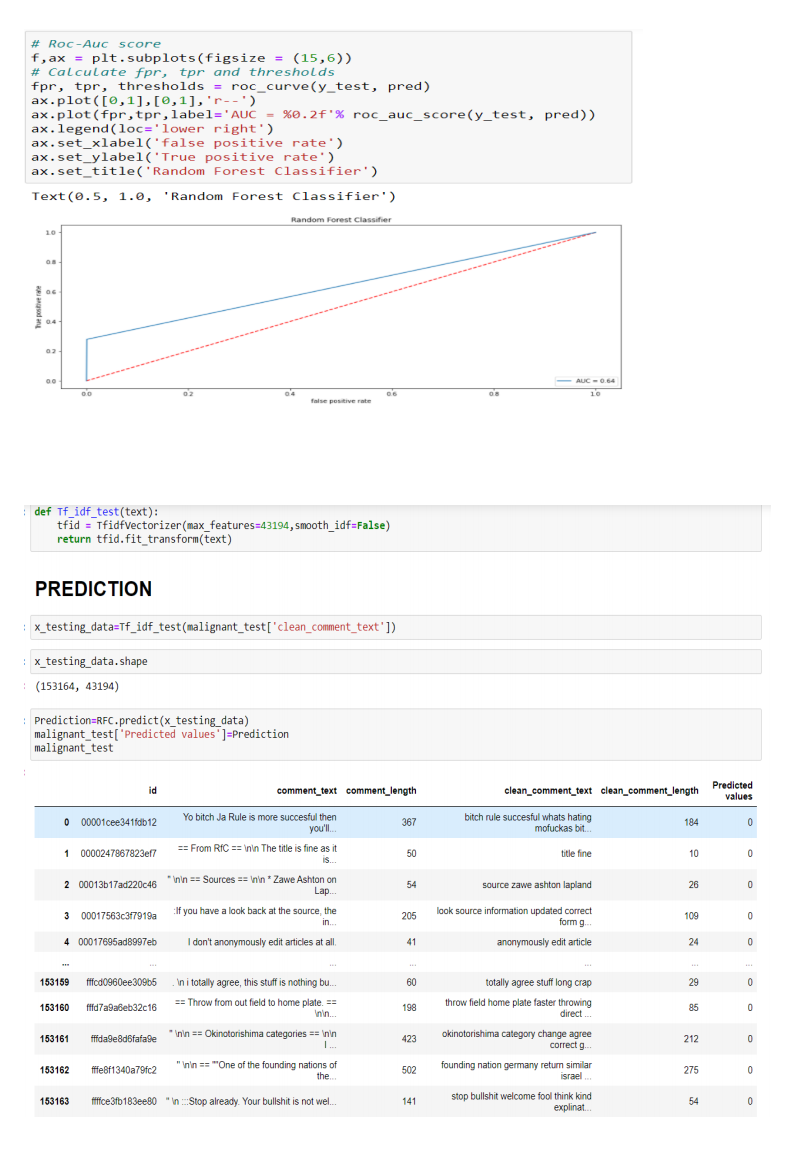


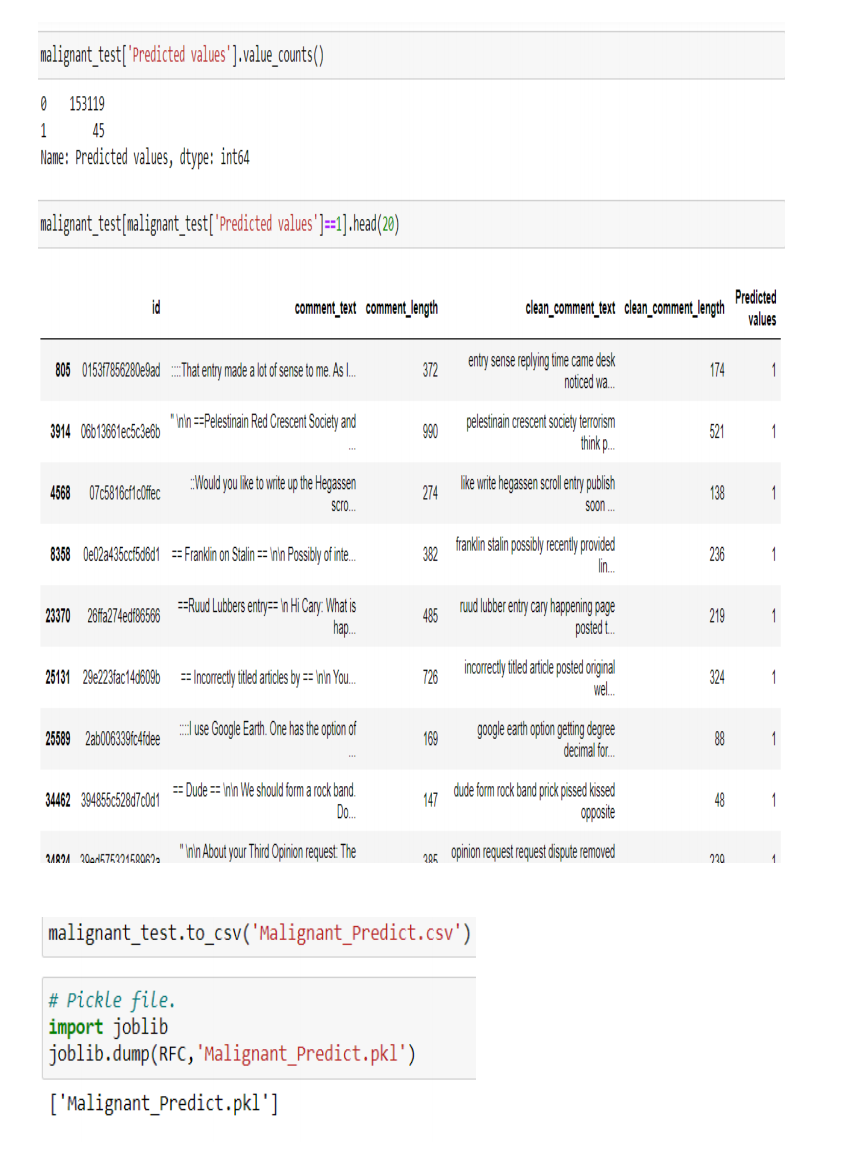


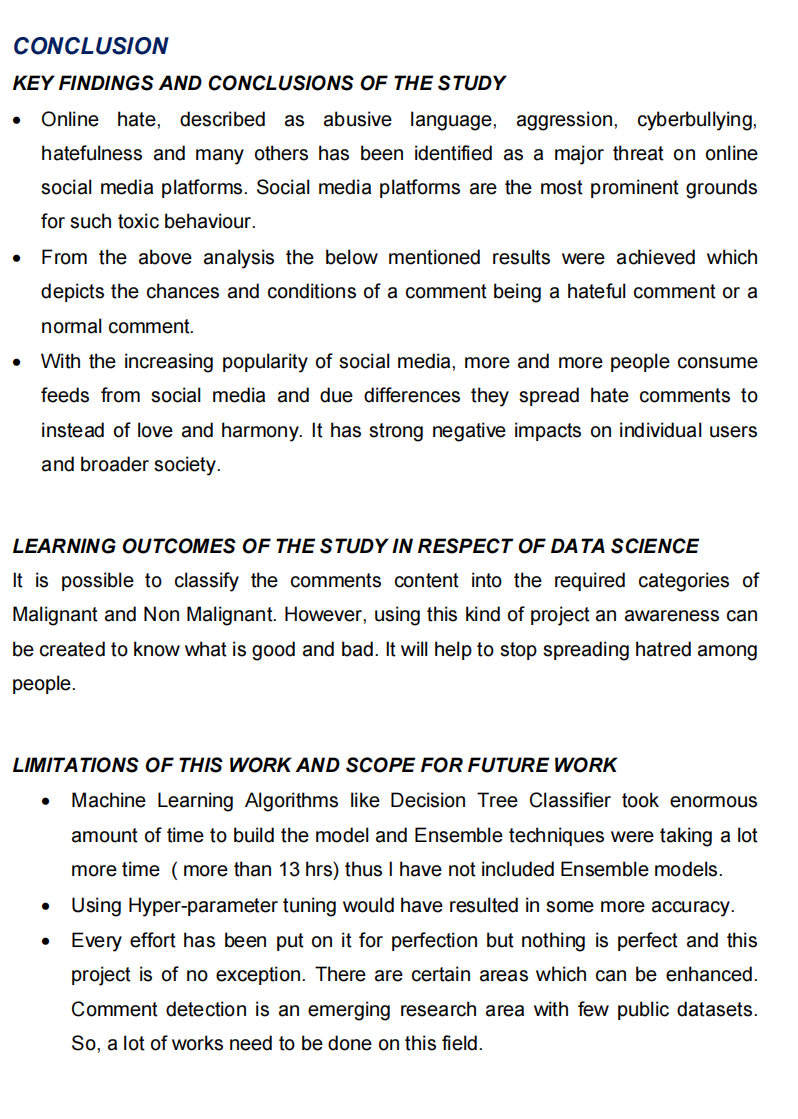












Thank You