

**MALIGNANT COMMENTS CLASSIFICATION**



**Submitted by:**

**ARUN KUMAR M**

**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, cyber bullying, 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 cyber bullying 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 inoffensive, 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 cyber bullying.

***DATA COLLECTION:***

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 0denotes 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.

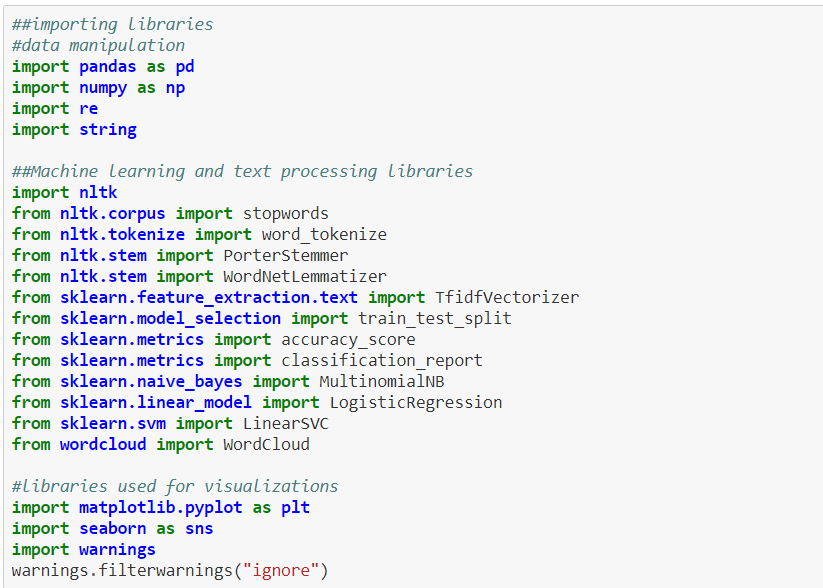
**Loathe:** It describes the comments which are hateful and loathing in nature.

**ID:** It includes unique Ids associated with each comment text given.

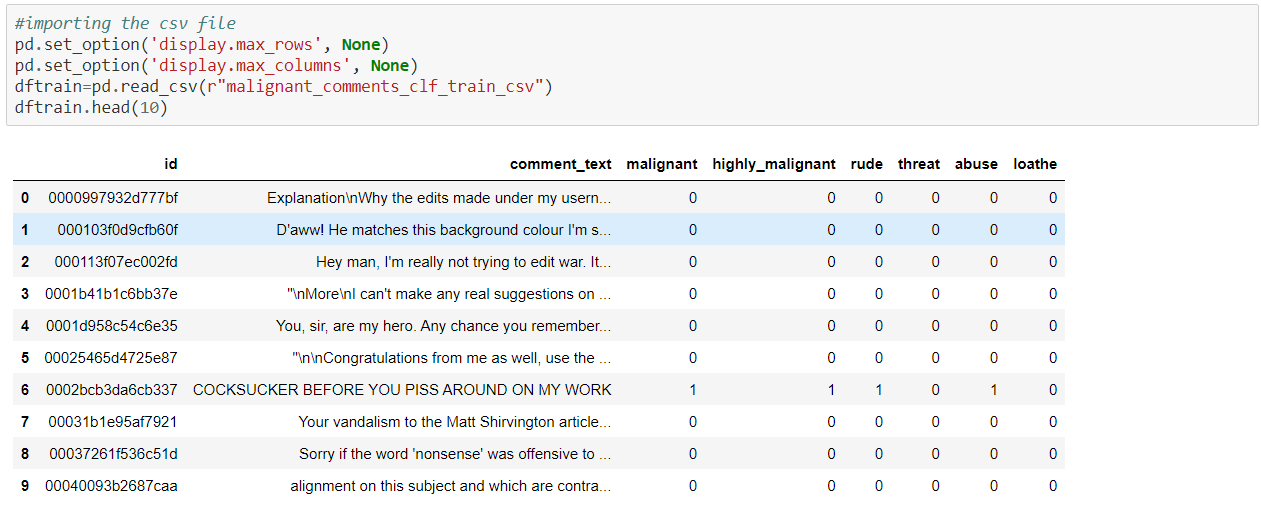
**Comment text:** This column contains the comments extracted from various social media platforms.

We could see that the data-set contains 115379 entries. Also, we can see that the data is imbalanced. Ratings counts differ for each rating.

***Importing Libraries:***



*Reading the csv file:*

**

***Basic statistical data:***

****

***Creating new column for comments length:***

******

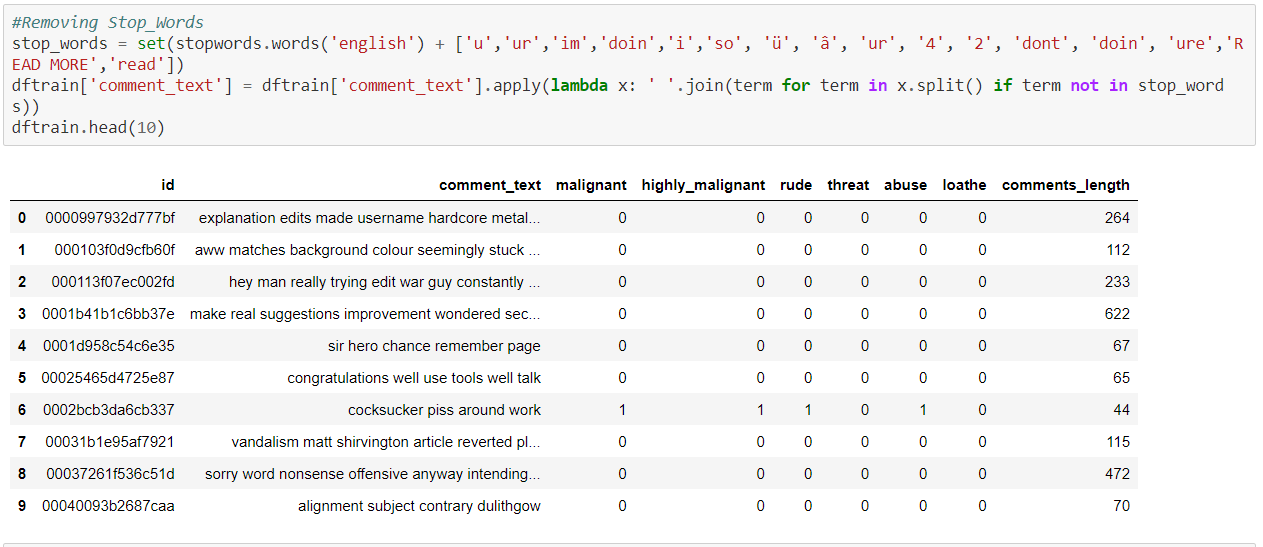
***Cleaning the data:***

*Removing the lower case, email address, web address, signs,*

*Phone number, numbers, punctuations, spaces, empty spaces.*

****

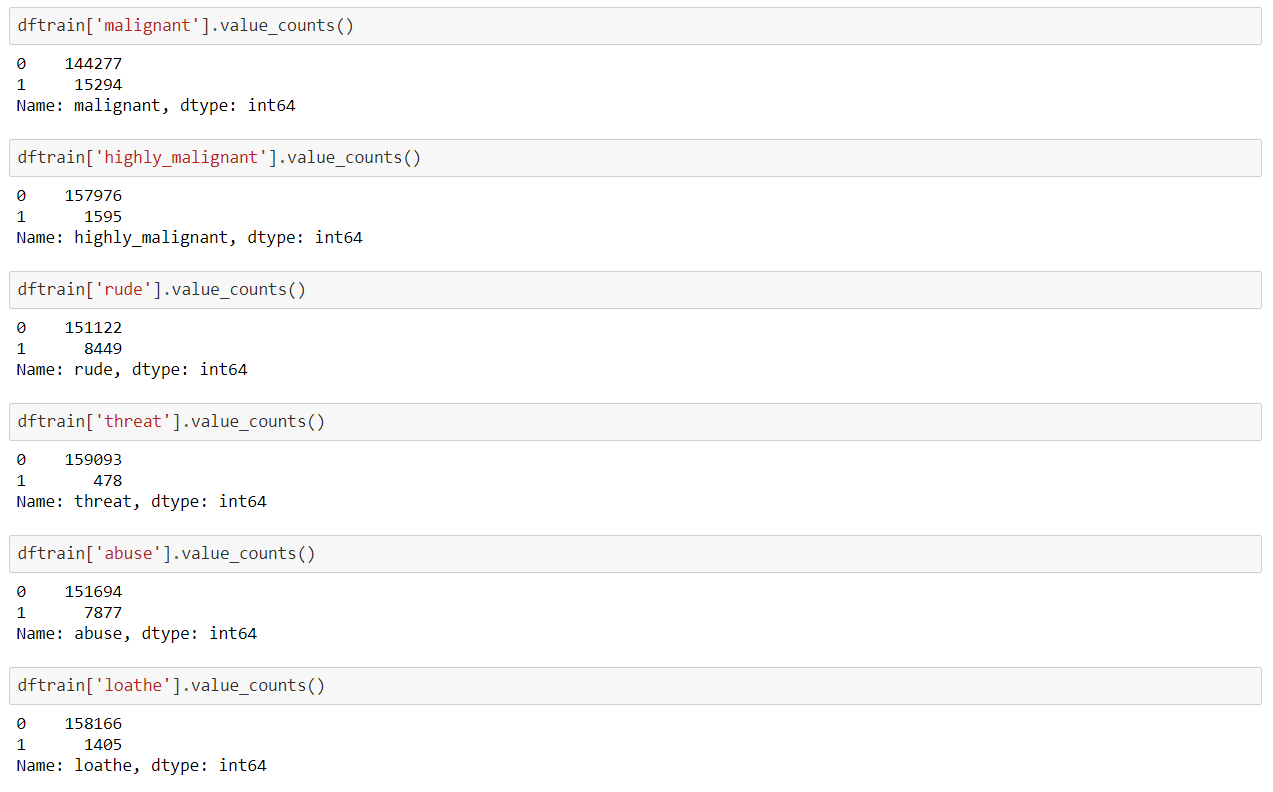
*Removing the stop words,*



***Creating a new column for cleaned comments length:***

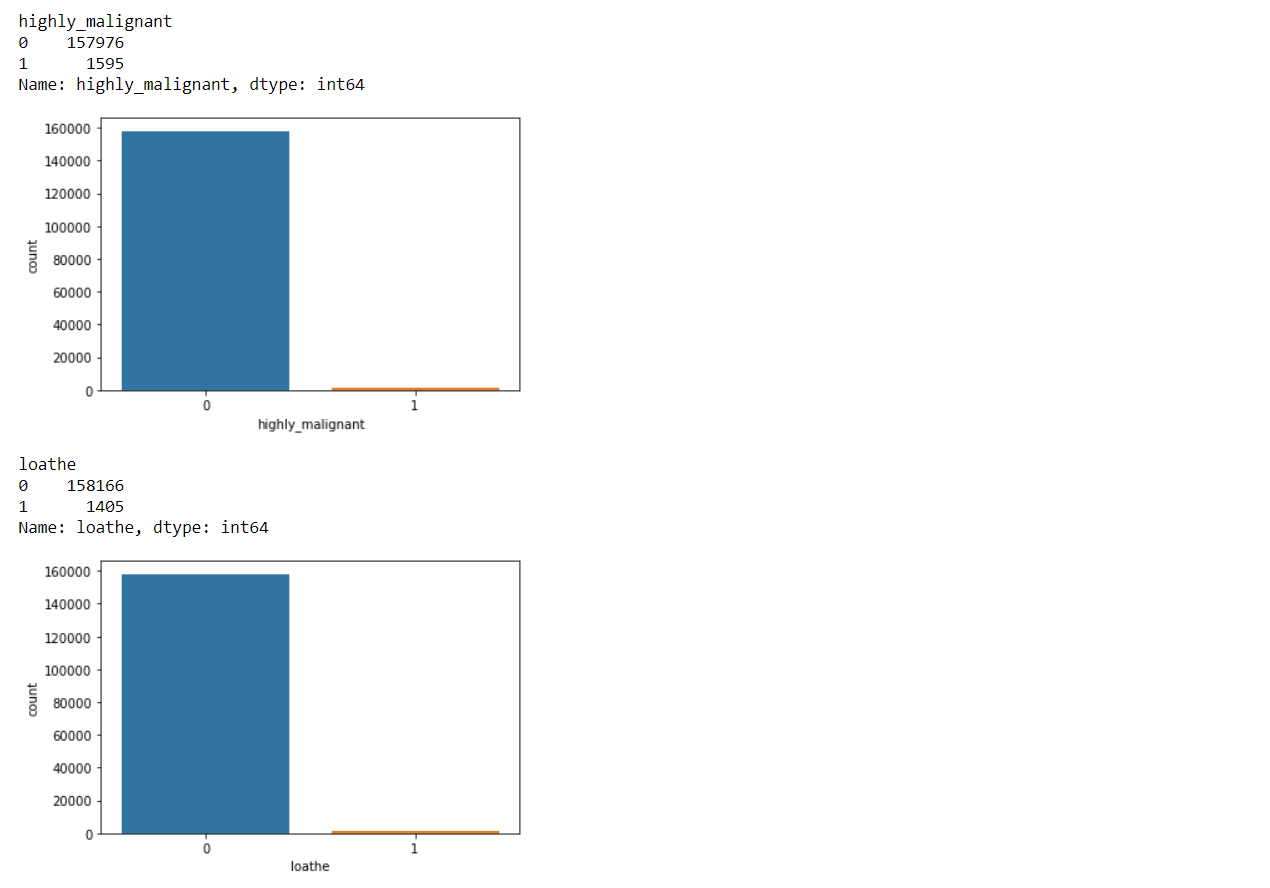
******

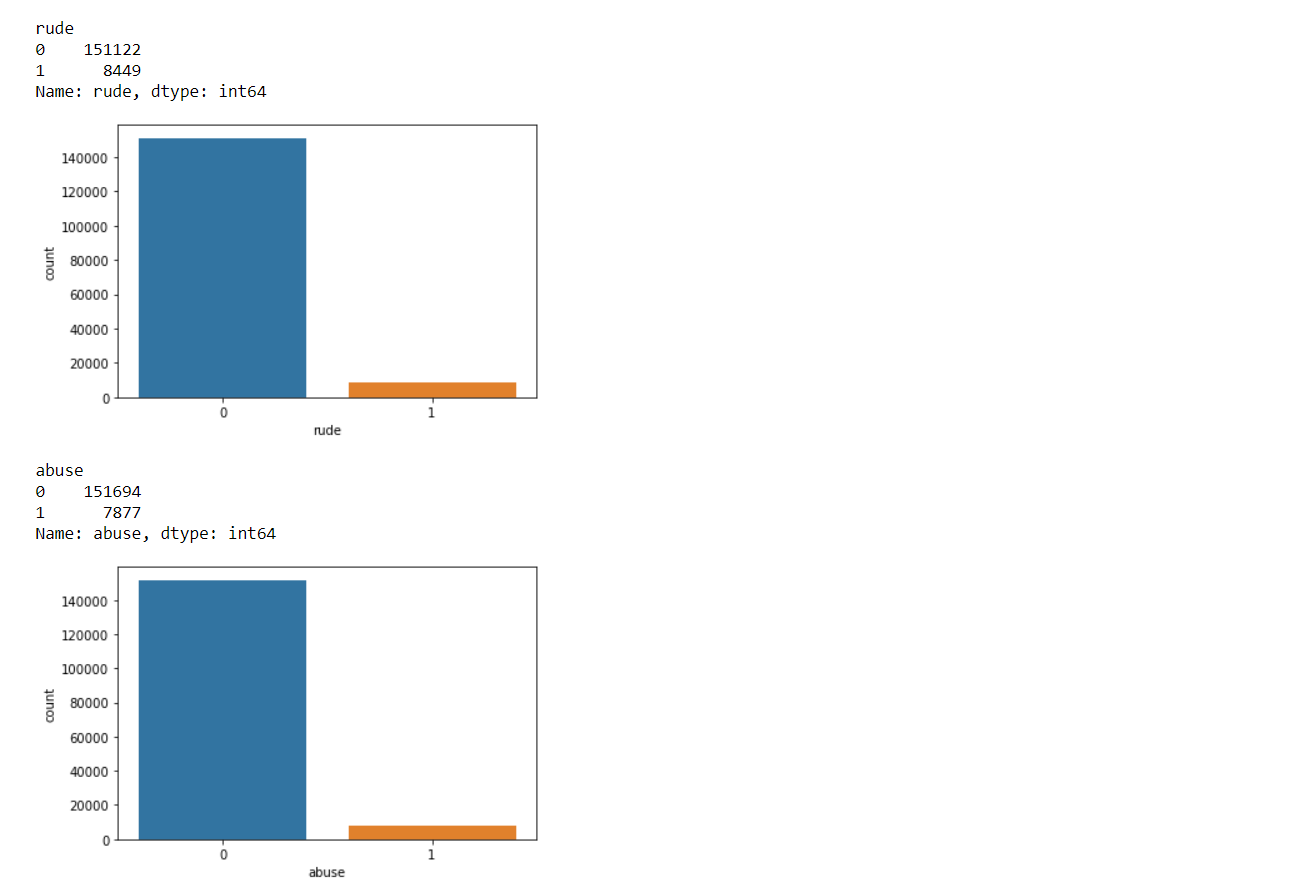
***Checking for target value counts:***

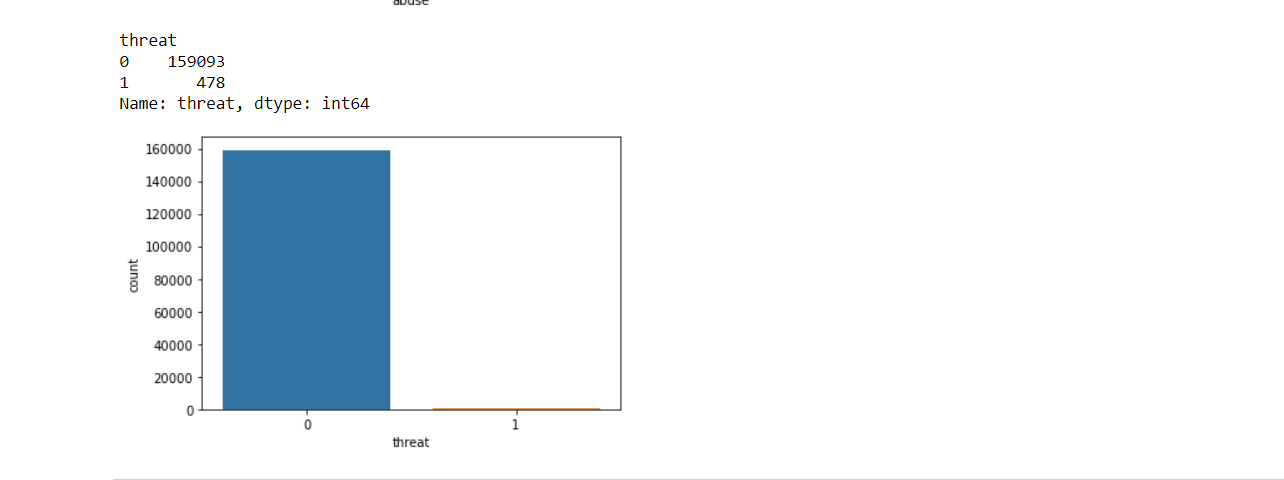
******

***Plotting for the target values:***

******

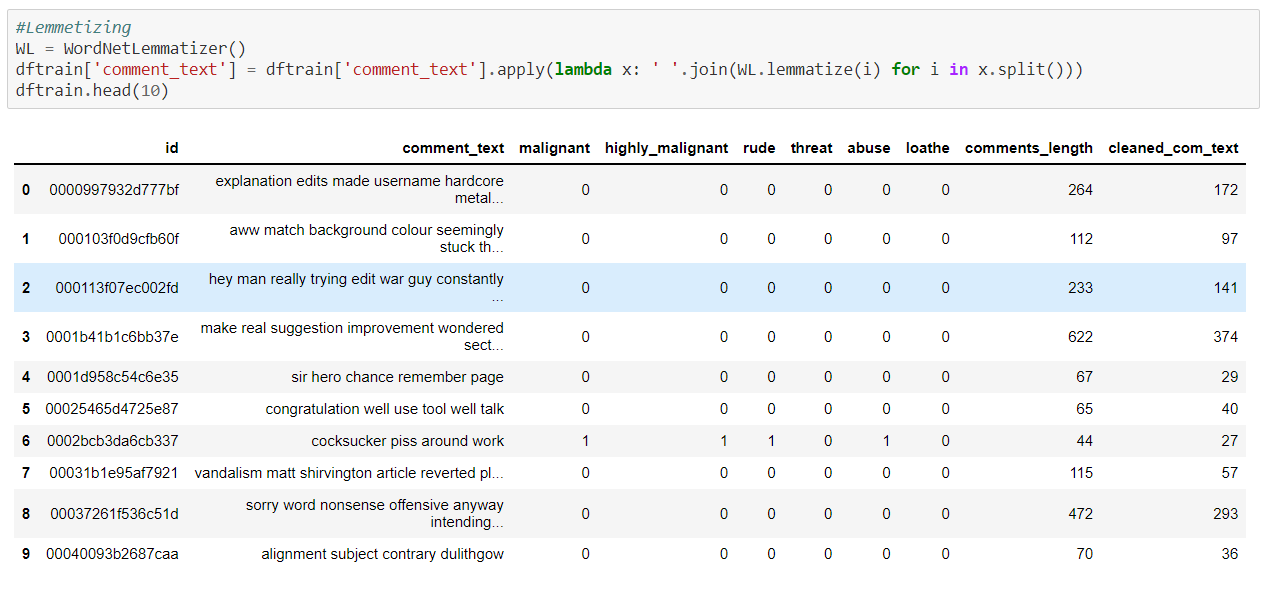
******

******

******

***Lemmatizing:***

*Lemmatizing the given comment text for data processing.*

******

***Getting the sense of words in the comment texts:***

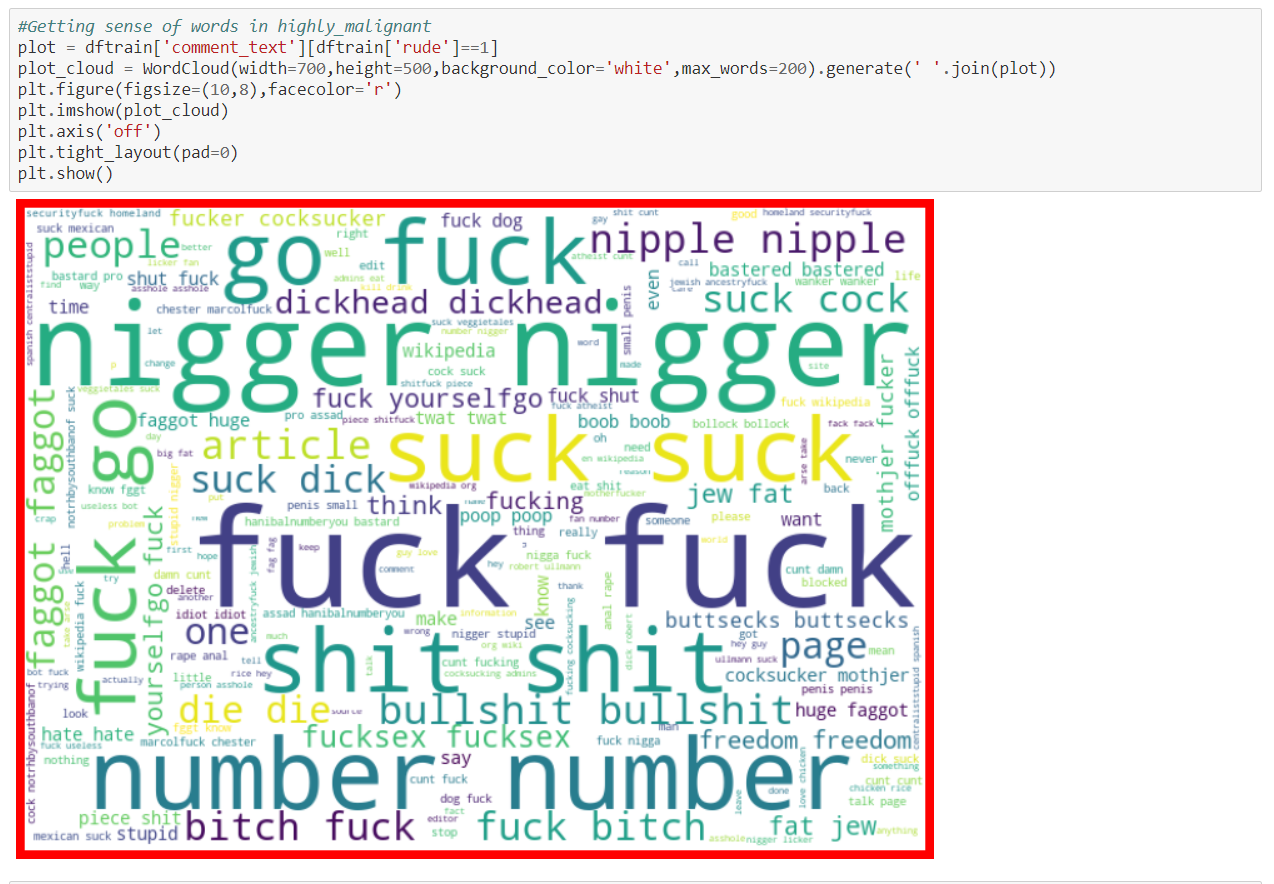
*Getting the sense of words in the malignant comments.*

******

*Getting the sense of words in the highly malignant comments.*

**

*Getting the sense of words in the rude comments.*

**

*Getting the sense of words in the threat comments.*

**

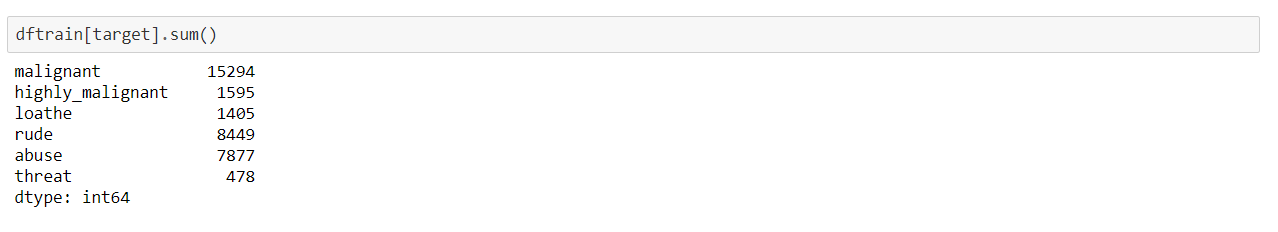
*Getting the sense of words in the abuse comments.*

**

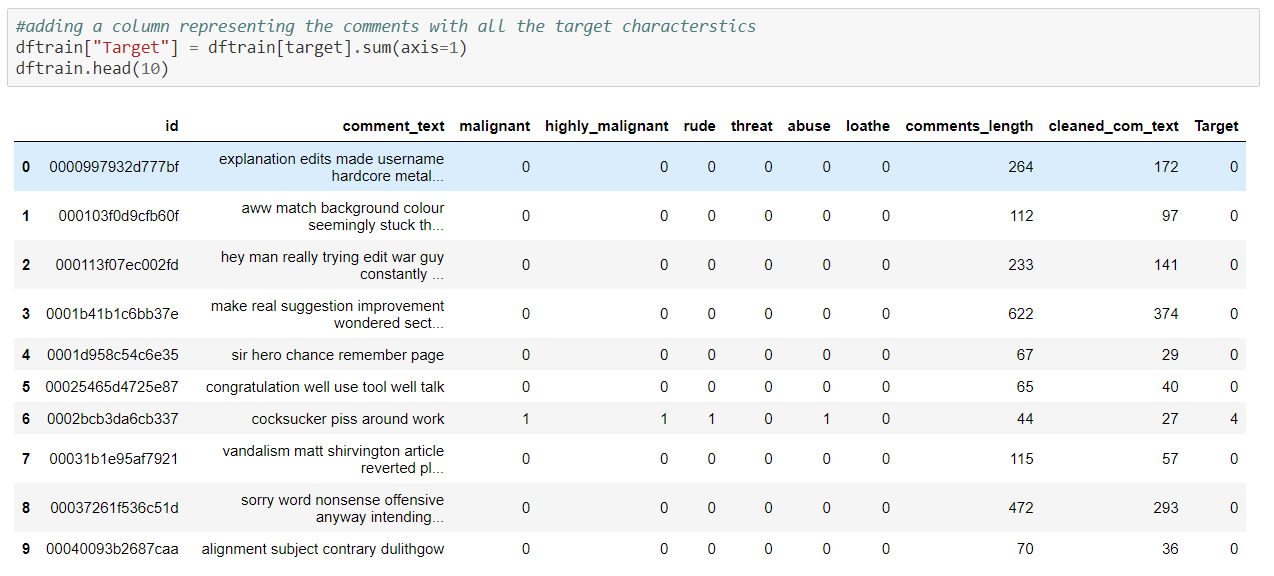
*Getting the sense of words in the loathe comments.*

**

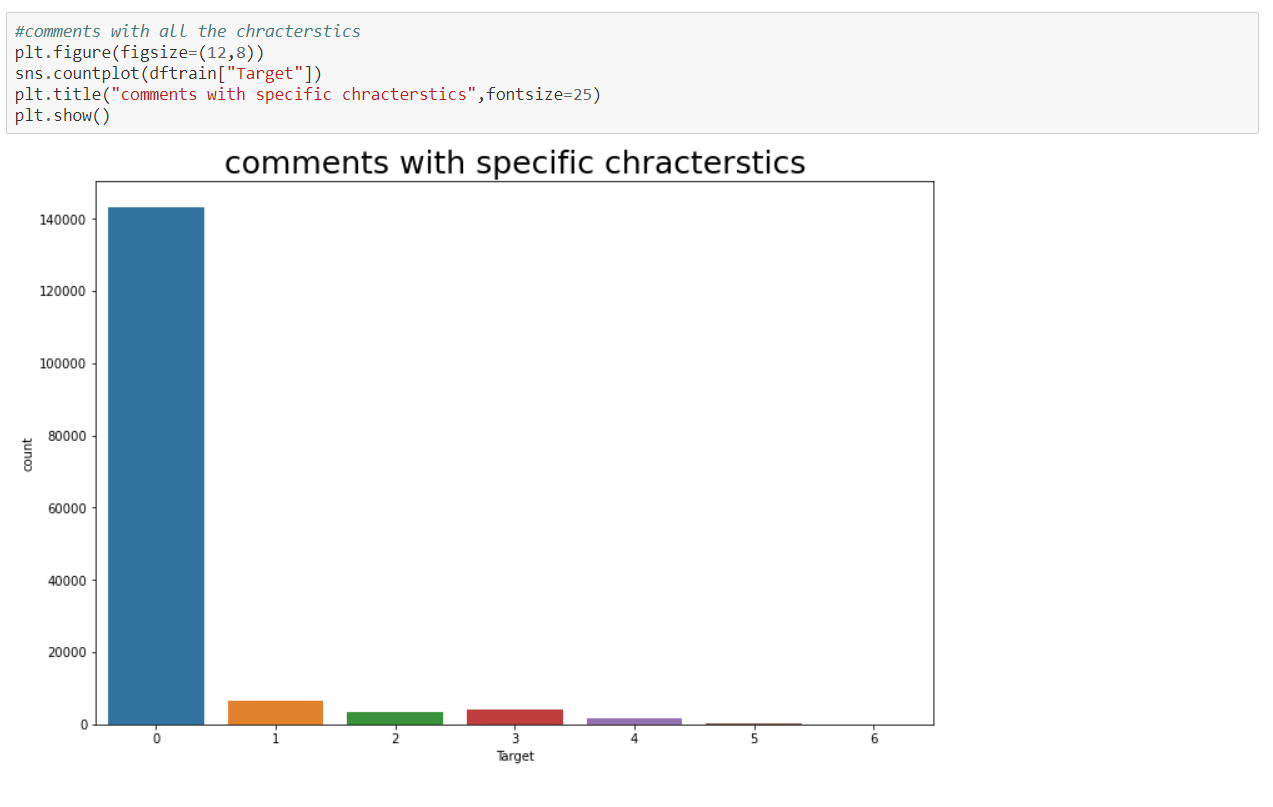
***Sum of target values:***

******

***Adding a column representing the comments with all the target characterstics:***

******

**

**

***Feature Extraction:***

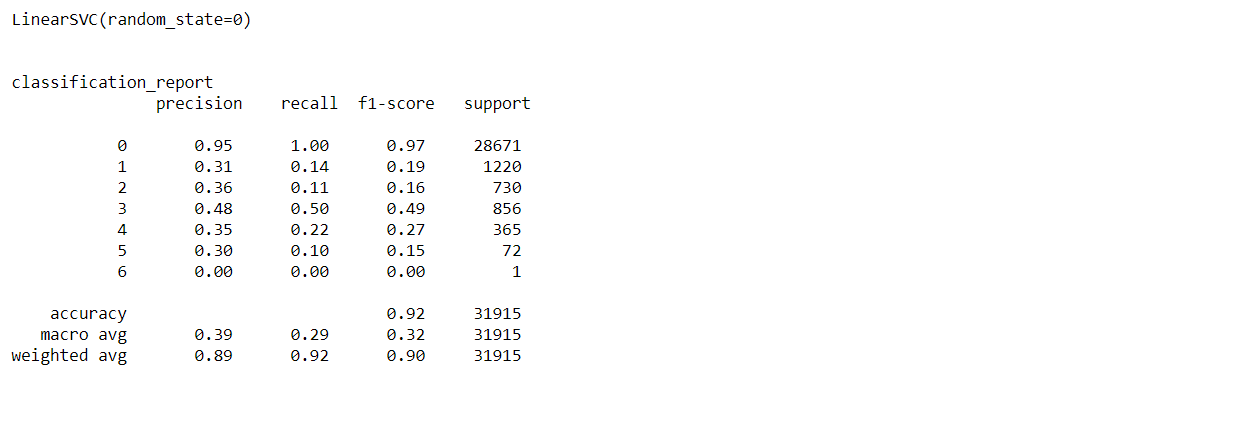
******

***Model Building:***

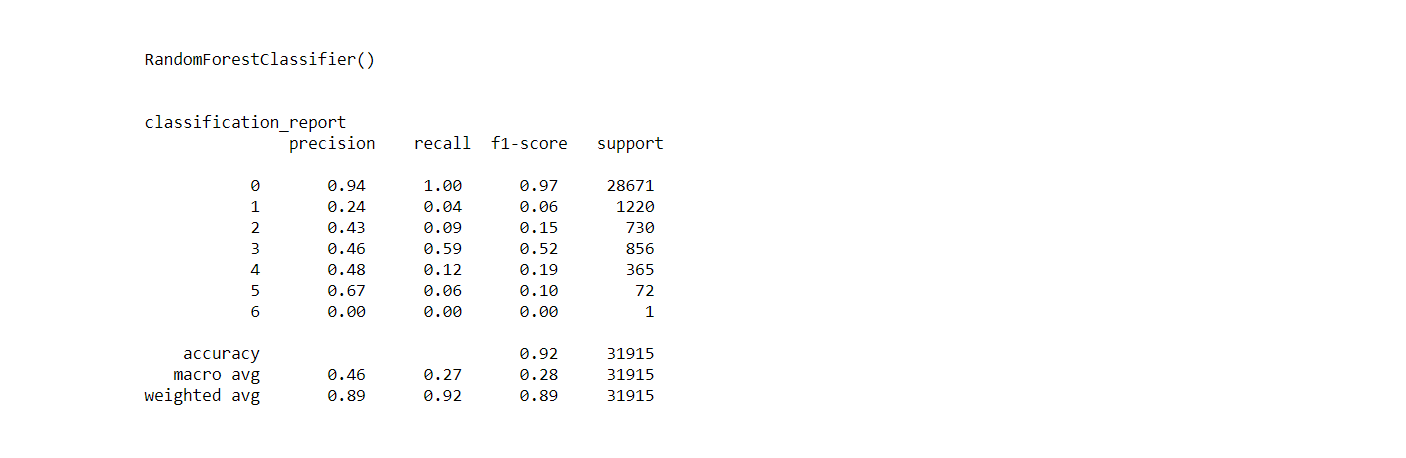
******

**

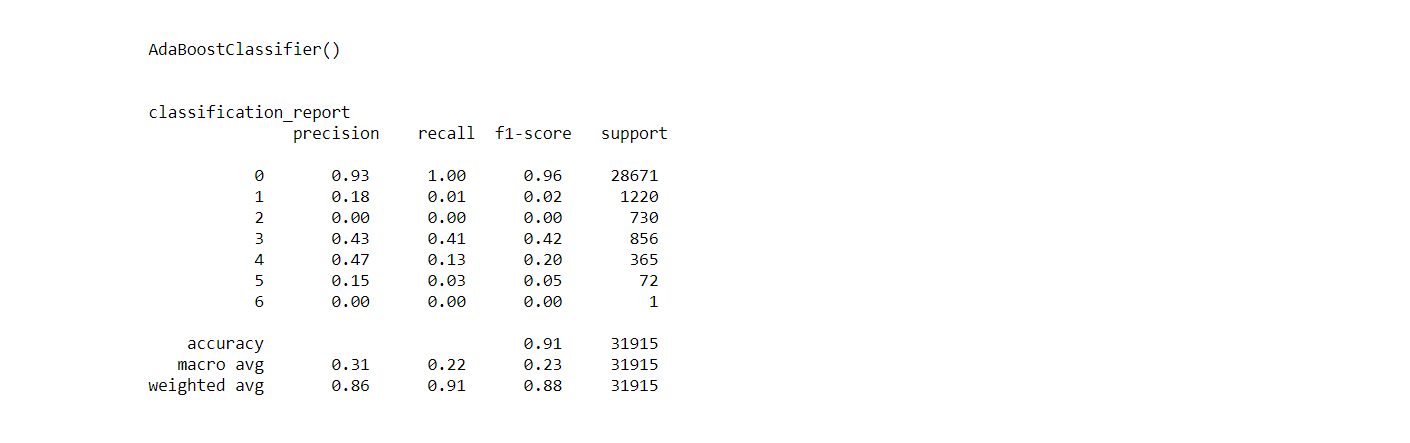
***LinearSVC:***

******

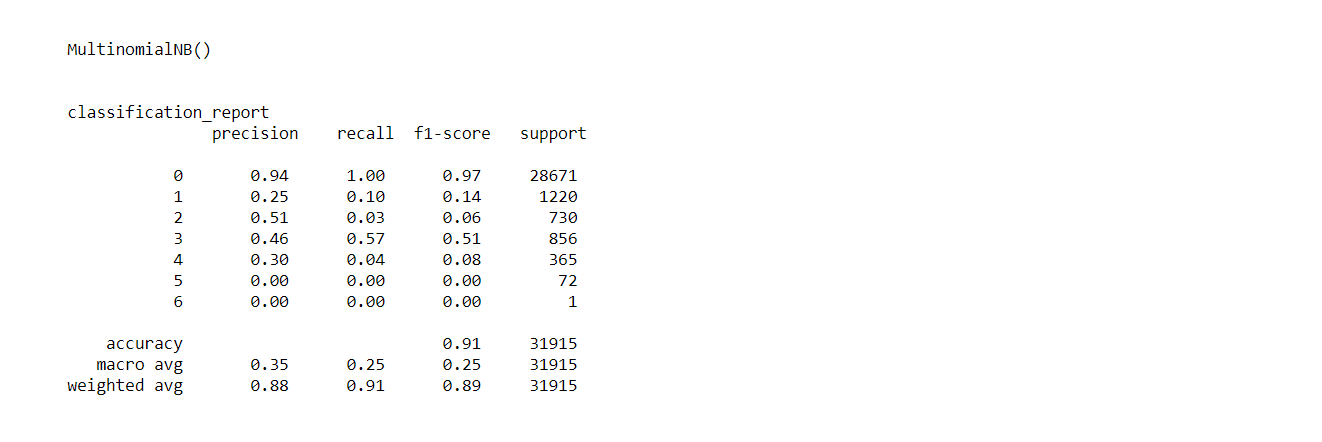
***RandomForestClassifier:***

******

***AdaBoostClassifier:***

******

***MultinomialDB:***

******