

***MALIGNANT COMMENTS CLASSIFICATION PROJECT***

**Submitted by: Manali Raj**

## Flip Robo Technologies

***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. I am also grateful to Mr Keshav for his constant guidance and support.

## Flip Robo Technologies

***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 behavior.

Therefore, 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

Internet is one of the important inventions and a large number of persons are its users. These persons use this for different purposes. There are different social media platforms that are accessible to these users. Any user can make a post or spread the news through these online platforms. These platforms do not verify the users or their posts. So some of the users try to spread online hate. 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.

Flip Robo Technologies 1 | P a g e

#### REVIEW OF LITERATURE

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 cyberbullying.

There is a difference between the traditional and very famous multi-class classification, and the one which we will be using, which is the multi-label classification. In a multi-class classification, each instance is classified into one of three or more classes, whereas, in a multi-label classification, multiple labels (such as – toxic, severe-toxic, obscene, threat, insult or identity- hate) are to be predicted for the same instance.

Multiple ways are there to approach this classification problem. It can be done using –

Multi-label methods which belong to the problem transformation category: Label Power Set (LP), Binary Relevance (BR), BR+, and classifier chain.

Base and adapted algorithms like: J48 (Decision Tree), Naïve Bayes, k-Nearest-Neighbor (KNN), SMO (Support Vector Machines), and, BP-MLL neural networks.

Further, out of the total dataset used for experimenting these algorithms, 70% was used for training and 30% was used for testing. Each testing dataset was labelled and thus for each algorithm using the predictions and labels, calculation of metric such as hamming-loss, accuracy and log-loss was done. The final results have been complied on the basis of values obtained by algorithmic models in hamming-loss and log-loss combined.

2

# ANALYTICAL PROBLEM FRAMING

### Dataset description

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.

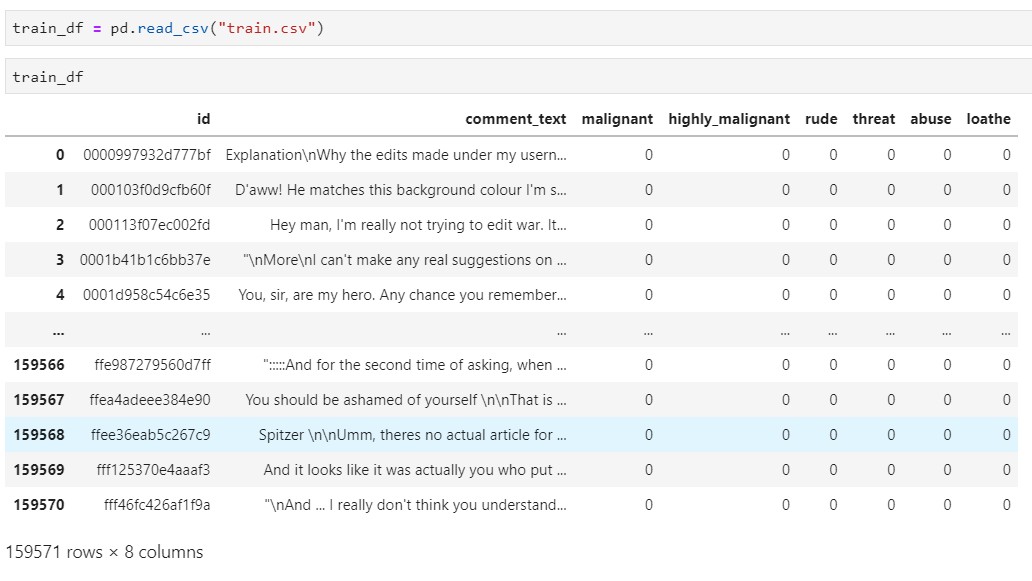
The data set includes:

* **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.

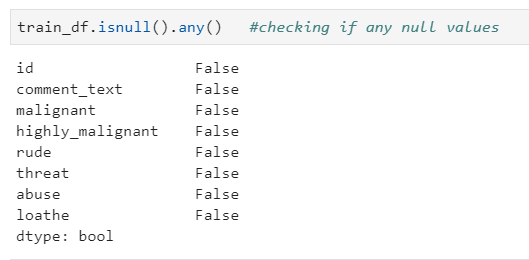
#### MODEL DEVELOPMENT AND EVALUATION

**Exploratory Data Analysis**

3

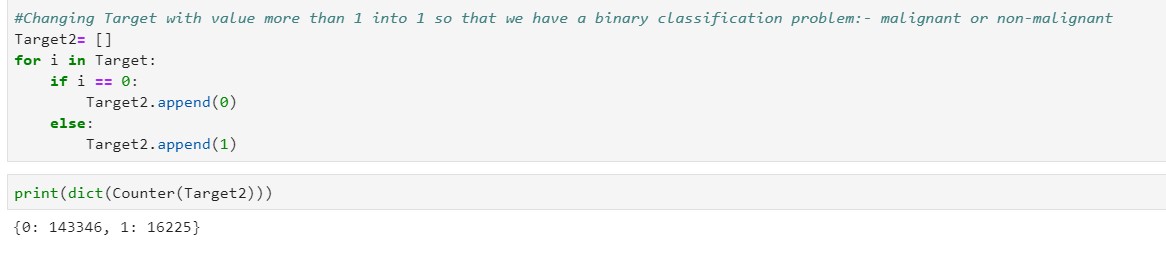
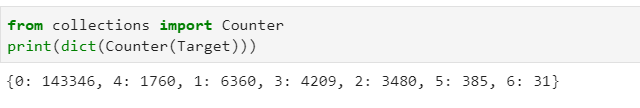


* + There are 159571 rows and 8 columns in the entire dataset.



* + Na values does not exist in the Dataset.

4



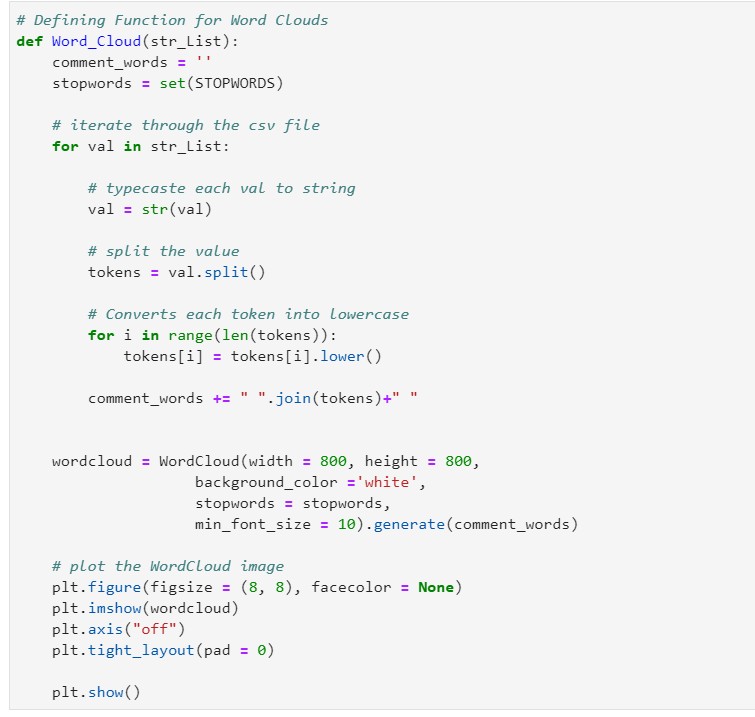
* Collecting all different types of outputs in a list

Here we can see that the dataset contains mostly outputs with 0 i.e. non-malignant comments and if Target > 1 indicating that the comment falls in more than one category such as rude, malignant, loathe, abuse etc.

Here Label "1" are offensive comments and "0" are non-offensive.

5

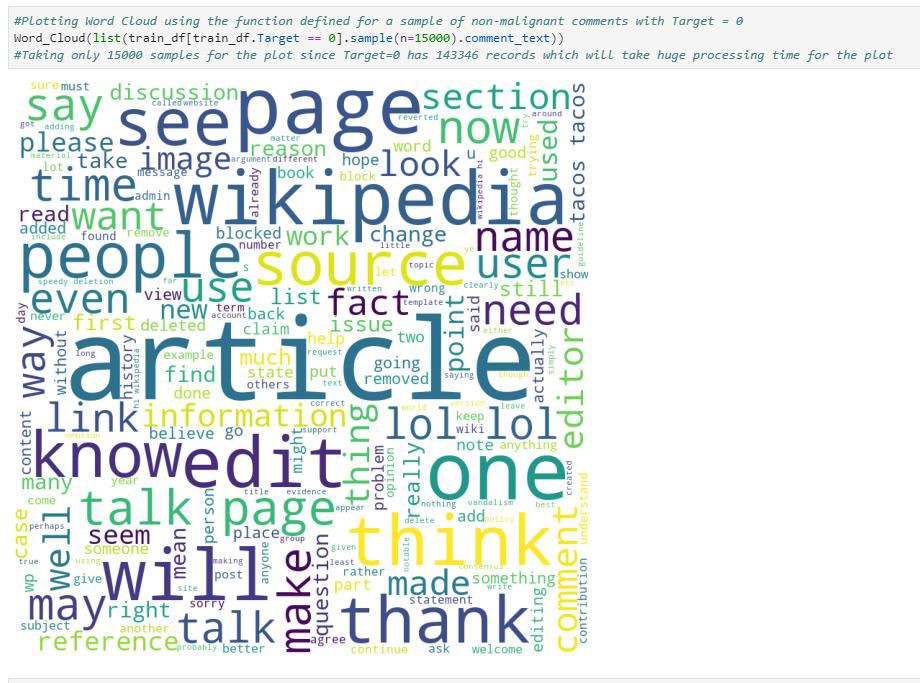
* Defining function for word clouds.





* Plotting word cloud function.

6



7

* Plotting word cloud function for malignant comments.



* Plotting word cloud for all types of malignant comments.

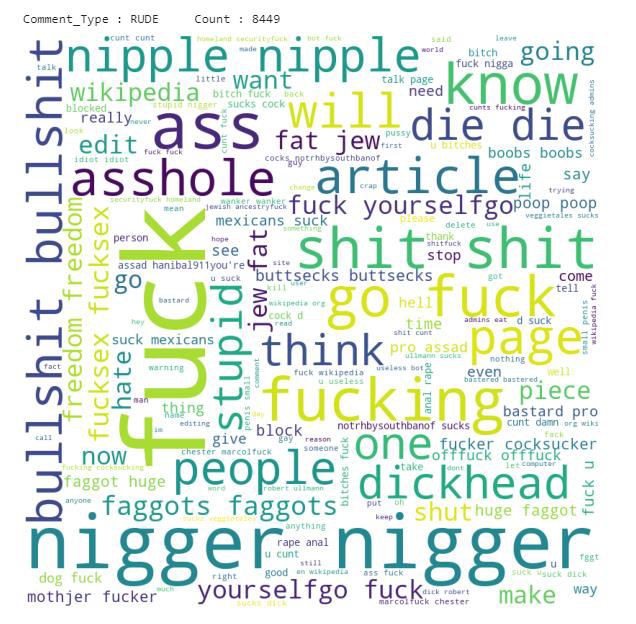
8



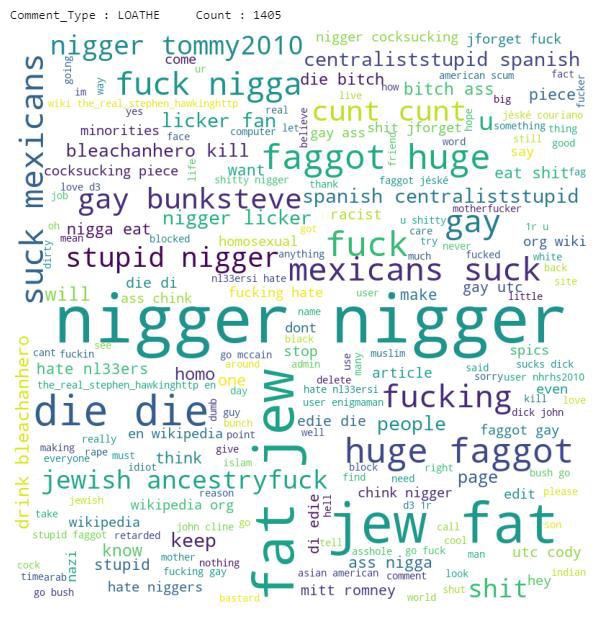
9



10



11



#### KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

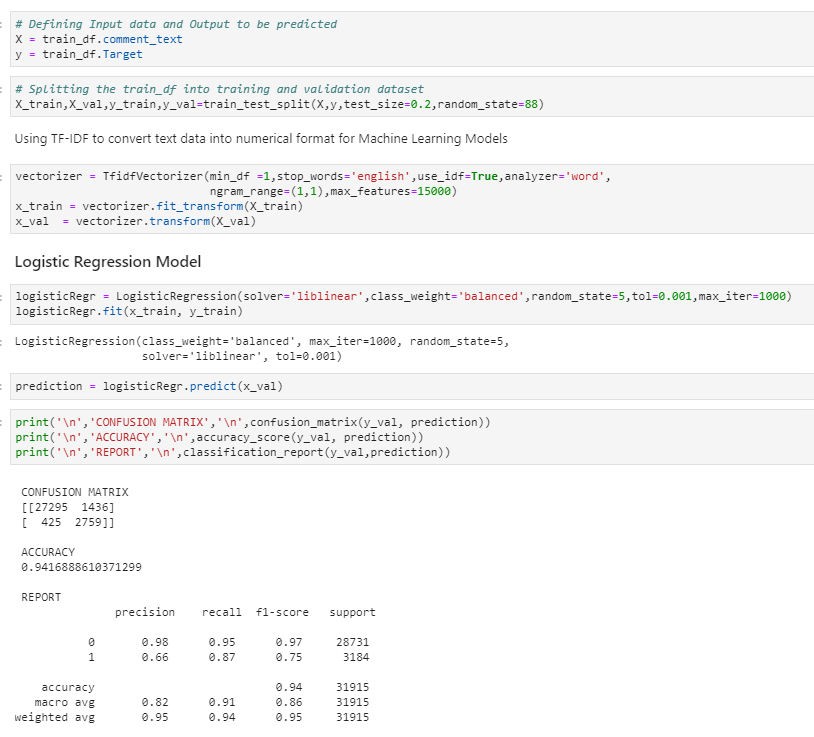
* + When it comes to the evaluation of a data science model’s performance,

sometimes accuracy may not be the best indicator.

* + So, we have used f1 score as well as recall, precision to check the performance of the models.

12

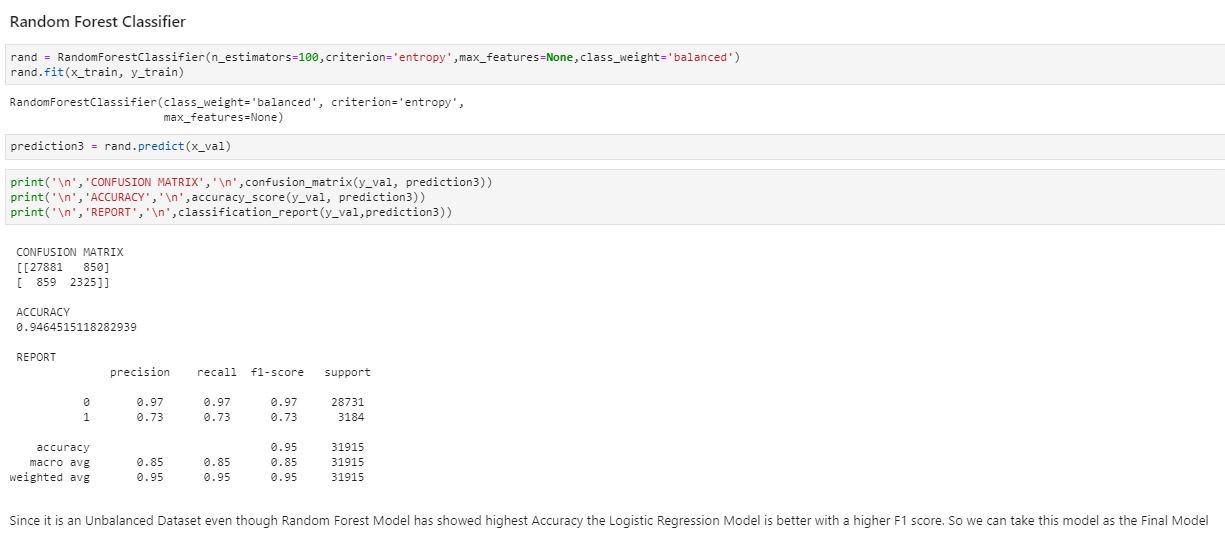
#### MODEL TRAINING



* + - Testing models for Malignant comments prediction

13

* + Four ML models were used to train the dataset using Sklearn, out of which the best was the Random Forest Model.



* Checking for the performance of the model

14

### CONCLUSION

* + Converted the problem into a binary classification problem. If the value in any of the columns: malignant, loathe, rude etc == 1 clubbed them together into one group and for rows with values == 0 counted them as non-malignent.
  + Plotted word cloud for all the different types: 'malignant', 'highly\_malignant', 'rude', 'threat','abuse', 'loathe'.
  + Logistic Regression was the best model with 94% accuracy and best f1 score.

15