Flip Robo Technologies

Malignant Comment Classifier Project

Submitted by: John Tojo

Data Science Intern at Flip Robo Technologies

ACKNOWLEDGMENT

It gives me immense pleasure to deliver this report. Working on this project was a great learning experience that helped me attain in-depth knowledge on data analysis process. Flip Robo Technologies (Bangalore) provided all of the necessary information and datasets, required for the completion of the project. I express my gratitude to my SME, [**Gulshana Chaudhary**](https://www.flipnwork.com/index.php/team_members/view/2282), for providing the dataset and directions for carrying out the case study procedure.

INTROUDCTION

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.

**Problem Statement**

* 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.

## 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:

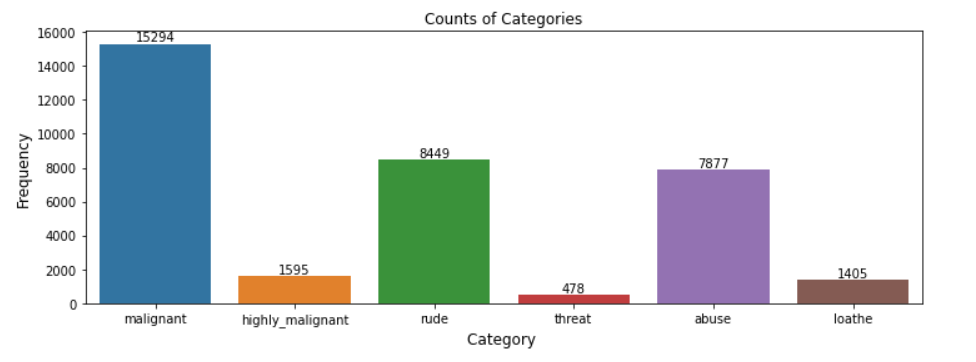
The description of each of the column is given below:

* **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.

**Exploratory Data Analysis (EDA)**

* Checked for nulls: no nulls were present
* Checked to data types: id and comment text were object datatype rest all were numerical datatype
* The number of positive and negative comments present were then stored into a data frame df\_fig to visualise the distribution of data provided
* It was observed that while trying to clean the text the loop would stop running after several iteration in order to overcome the recursion limit was raised from 3000 to 10000.
* The clean text was stored in corpus and stored into a data frame df\_x which had all details, for comparison and for visualisation
* A table was then used to see the difference in most frequent used words before and after cleaning for positive and negative comments, which was later plotted using wordcloud as well
* Bag of words was used to convert the text to vectors
* The data was divided using train test split method, with random state 0 and test size as 20%

Visualization



it can be seen that among negative comments the most common type was malignant followed by rude

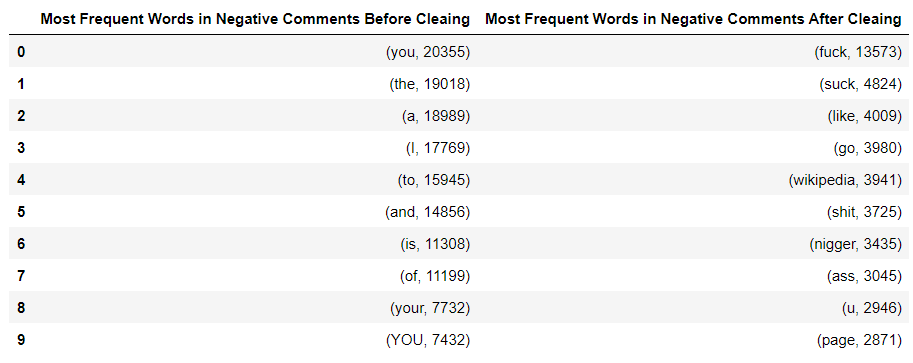


Table1: showing the frequency of negative words before and after cleaning

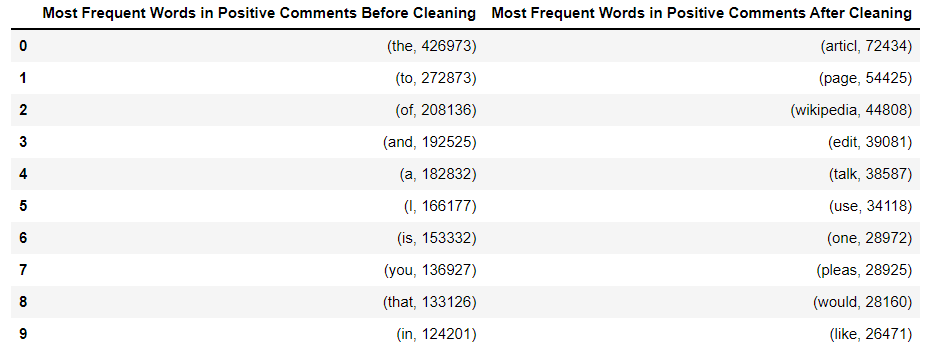


Table2: showing the frequency of positive words before and after cleaning

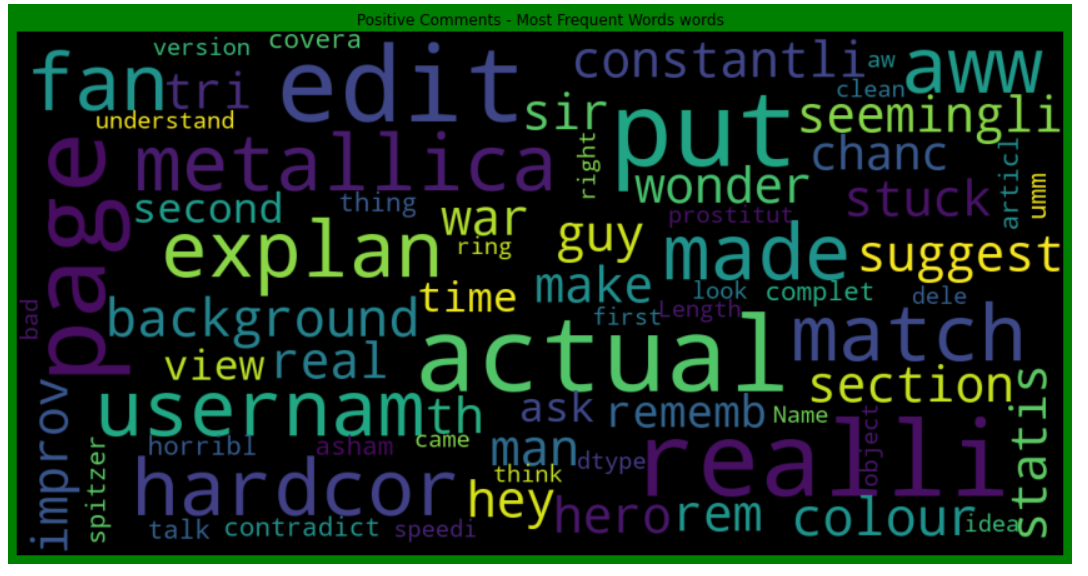


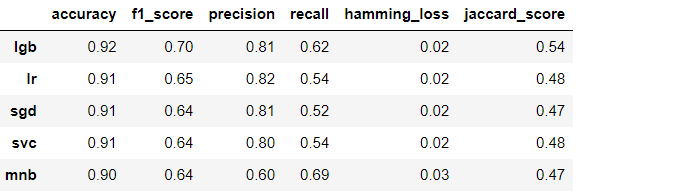
Fig1: showing wordcloud for positive comments



Fig2: showing wordcloud for positive comments

**Model Selection:**

* The problem was classification problem as the target variables would take value 0 or 1
* There were 6 target variables, OneVsRestClassifier was used
* The different models used were as follows:
  + LinearSVC
  + LogisticRegression(solver='lbfgs')
  + MultinomialNB
  + LGBMClassifier
  + SGDClassifier
* The paramters used to determine the best models were as follows:
  + accuracy
  + precision
  + recall
  + jaccard score
  + hamming\_loss



It can be seen that the best model was lgb

1. highest accuracy, precision, recall, jaccard score
2. least hamming\_loss

**Predictions**

The data on which the prediction were to be done underwent the same process as train dataset

The predictions were made and then stored into a dataframe x\_test



**Conclusion**

1. machine learning algorithms to classify the comments
2. visualization has helped in understanding most frequenlty used words for positive and negative comments