Malignant Comments Classifier

Submitted by:

Aravind S

ACKNOWLEDGMENT

I thank Data Trained team and the faculty at FlipRobo to get an exploration on a real-life dataset which helped in exploration of the concepts that I use. I have also taken few references from Kaggle and also some YouTube videos to find apt resolution for the dataset.

Business Problem:

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 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.

We will be removing the id columns later as that does not give a relevance to the output.

These are the outputs that we have:

- **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.

Just to check the comment length, we had checked the length using str.

We had also used a function to remove all the email address, removing punctuation, url's, Html tags, web addresses, word tokenization and removing the stop words.

Then comes the main part, that is cleaning the data.

Had made a separate column to check the lines were in no malignant words are there and findings were interesting.

Out for the total there are only 15294 data where Malignant data is present.

Out for the total there are only 1595 data where Highly Malignant data is present.

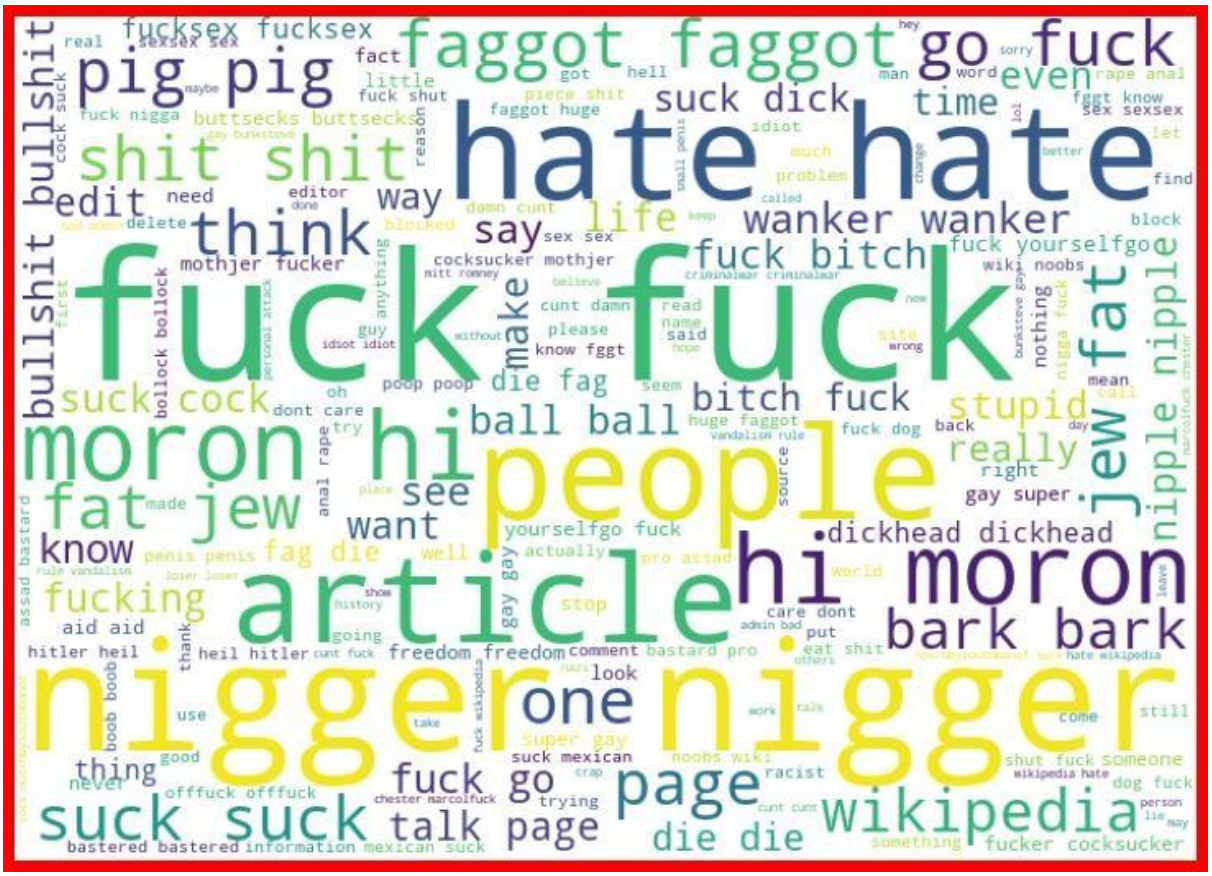
Out for the total there are only 1505 data where Loathe data is present.

Out for the total there are only 8449 data where Rude data is present.

Out for the total there are only 7877 data where Abuse data is present.

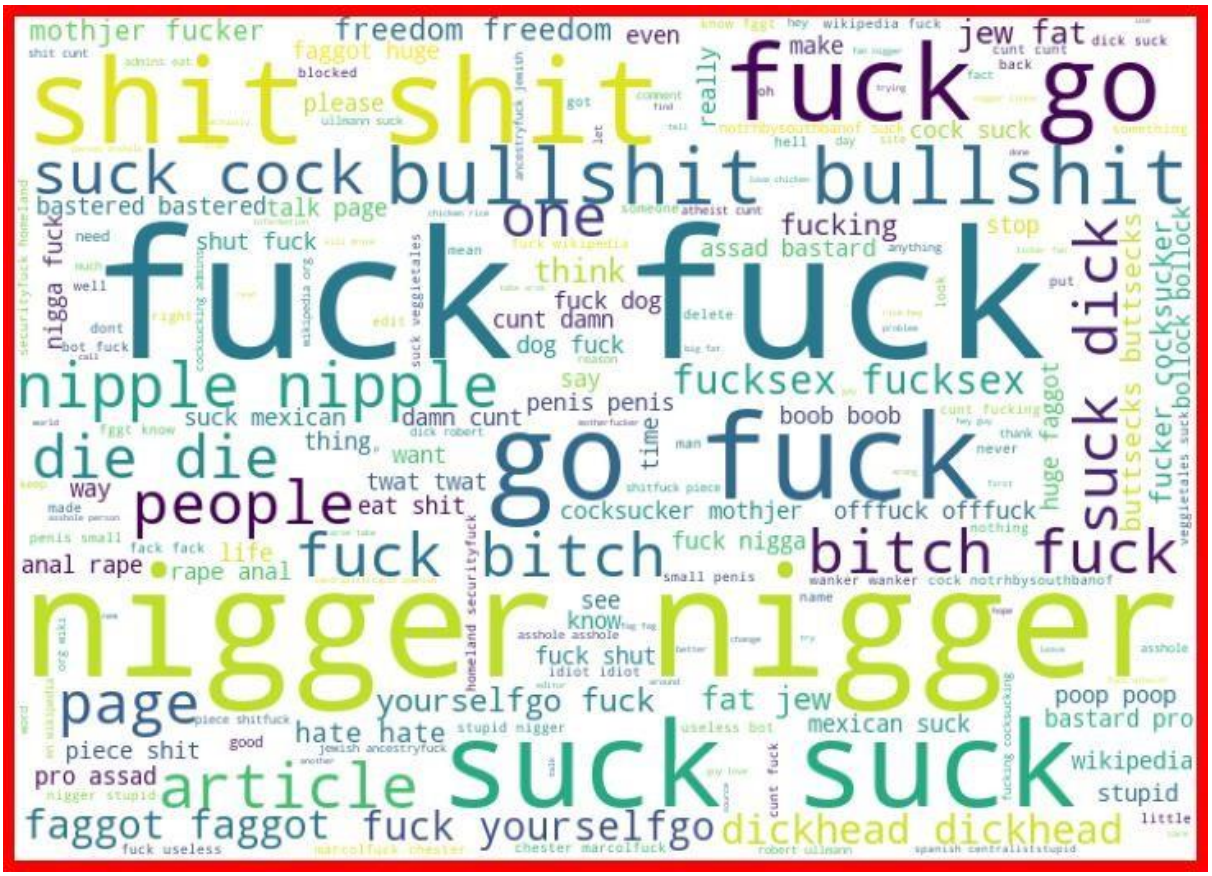
Out for the total there are only 478 data where Threat data is present.

Here is the image of Malignant data:



Here is the image of Highly Malignant:

Here is the image of Rude:



Here is the image for Threat:

Here is the image of Abuse:

We had dropped the ID column here.

Next procedure will be to vectorize the data to get the strings into numbers so that we can check the details to predict further, that will tokenize the documents. Using this system learns vocabulary and inverse document frequency weightings and allow you to encode new documents.

Then we can split the data into test and train here and as it is a multiclass classification model, we have to use multi learn classification in the model. Also we will check the linear regression and Navies bayes classification to predict the accuracy of the models.

We have to use estimators as well because in normal models do not work good on multi class classification and it cab be bypassed using estimators like Binary relevance and Classifier chain.

After running the model, we had got an accuracy of 91.99% with a log loss of 1.252 which works pretty fine.

Conclusion

In the provided data, we had tried a few classification models and also checked on how to work with vectorization to predict the accuracy of the test and the output was pretty good. Though we were having a logarithmic loss, that fluctuation was because of the possible imbalance in the data that we have, so taking that into consideration, we have to afford the loss here or we can use Smote to over sample the data, under sample will not be possible here as in NLP, data is the king and we cannot afford to lose data. After the prediction, we were able to get the accuracy of 91.99% over thetest data.