FLIP ROBO

DATA ANALYSIS REPORT ON

MALIGNANT COMMENT CLASSIFIER

Submitted by
Tinu Shiby
Data Science Intern
Flip Robo Technologies

ACKNOWLEDGMENT

I wish to express my sincere gratitude to my mentor Mr. Keshav Bensal, Flip Robo Technologies for his inspiration and guidance at every stage of the project. The blessing, help and guidance given by him will certainly help me in achieving better things in future.

I wish to express my profound gratitude to the DataTrained family for providing an opportunity to undertake this internship.

Lastly, I thank the almighty and my parents for their every small support and encouragement which helped me to complete this project successfully.

INTRODUCTION

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 an online hate and abuse comment classifier which can be used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

ABOUT THE DATASET

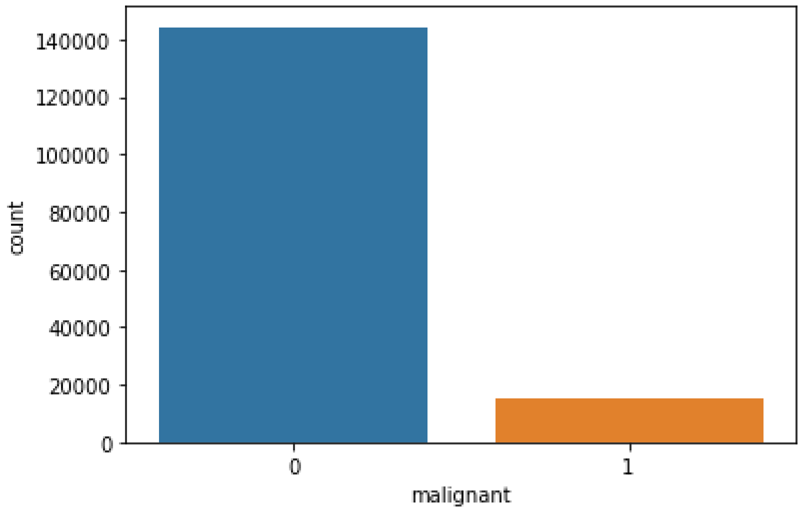
The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which include ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’. The label can be either 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 indications 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. This project is more about exploration, feature engineering and classification that can be done on this data.

EXPLORATORY DATA ANALYSIS

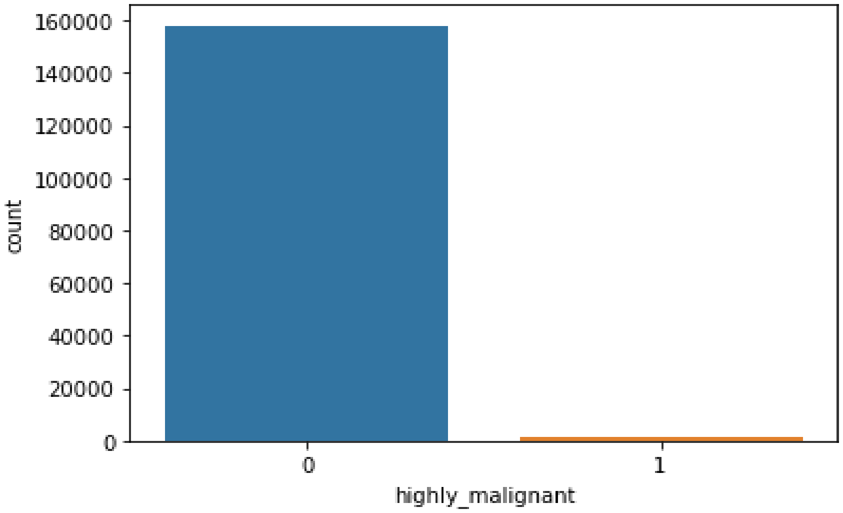
Plotting the count plot of all the variables

i) Malignant

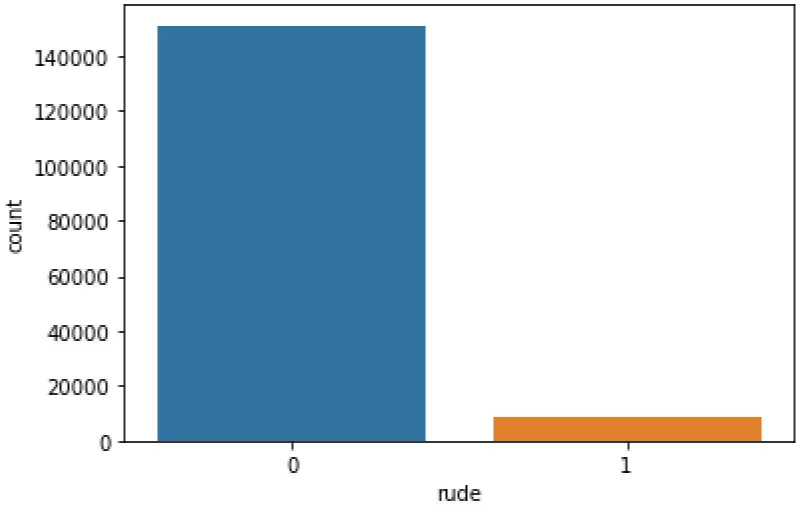


In the dataset the count of malignant message is comparatively low.

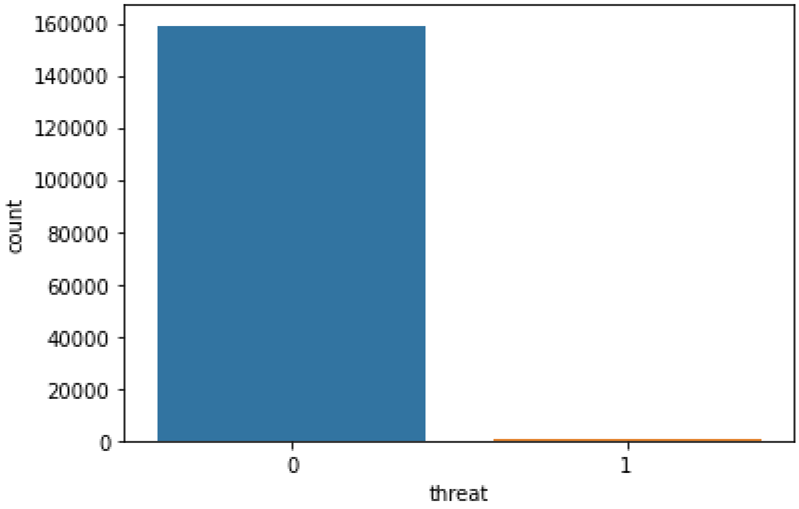
ii) Highly Malignant



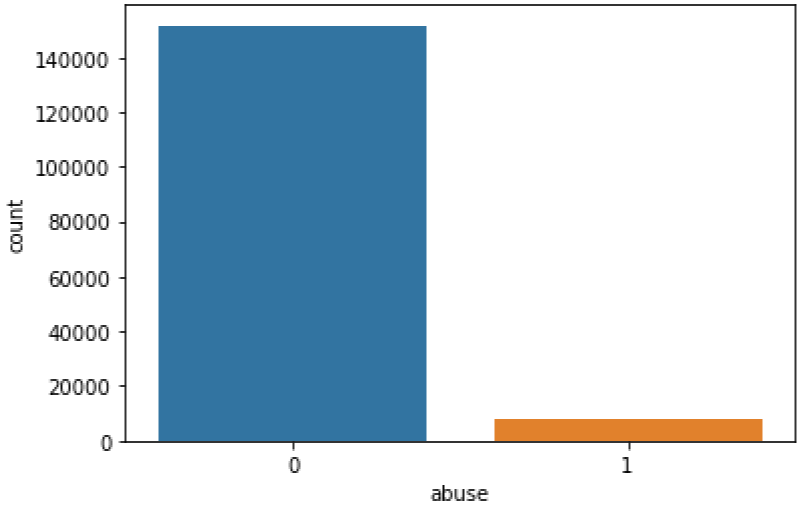
iii) Rude



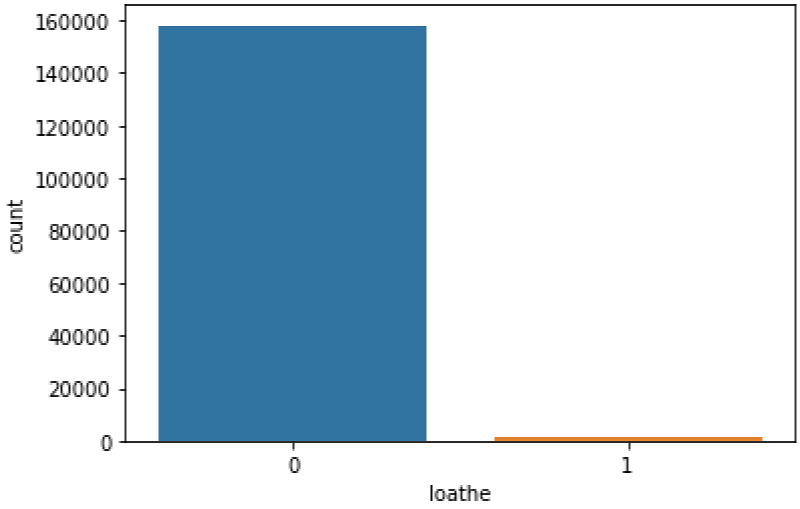
iv) Threat



v) Abuse

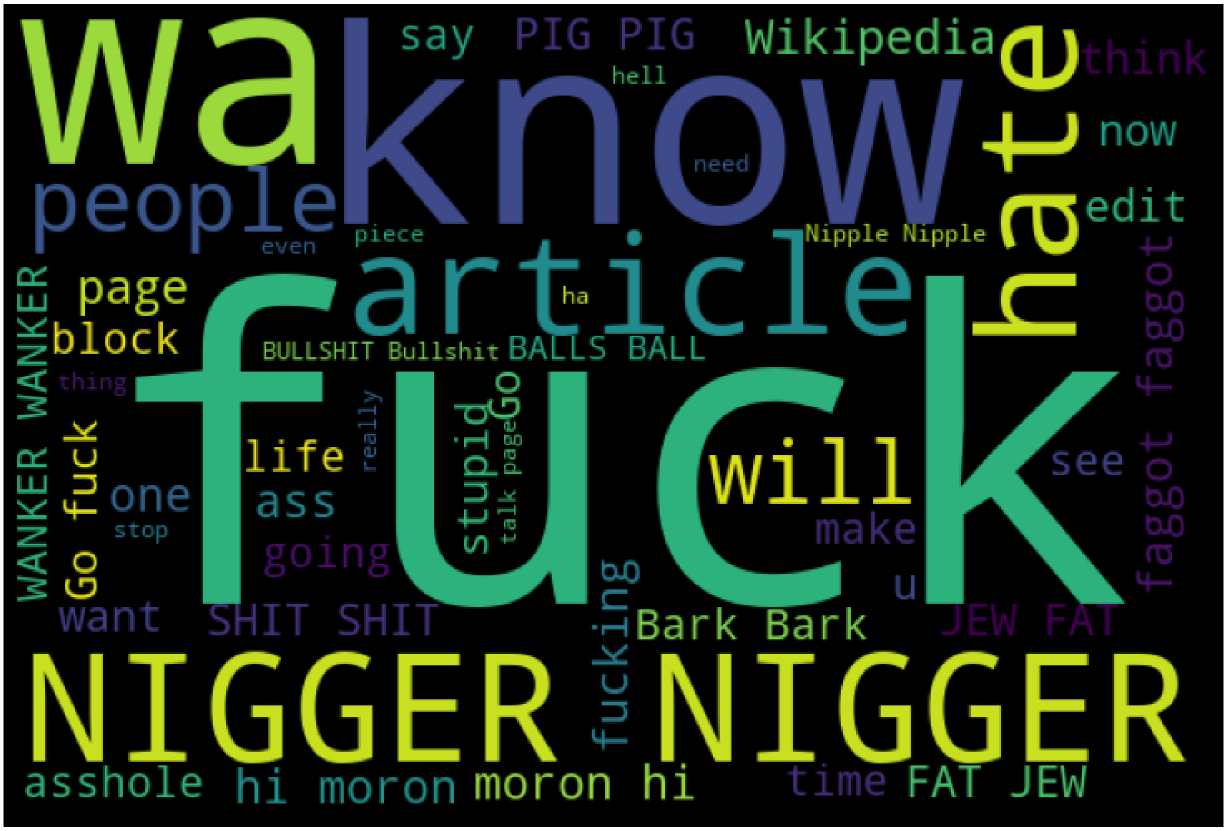


vi) Loathe

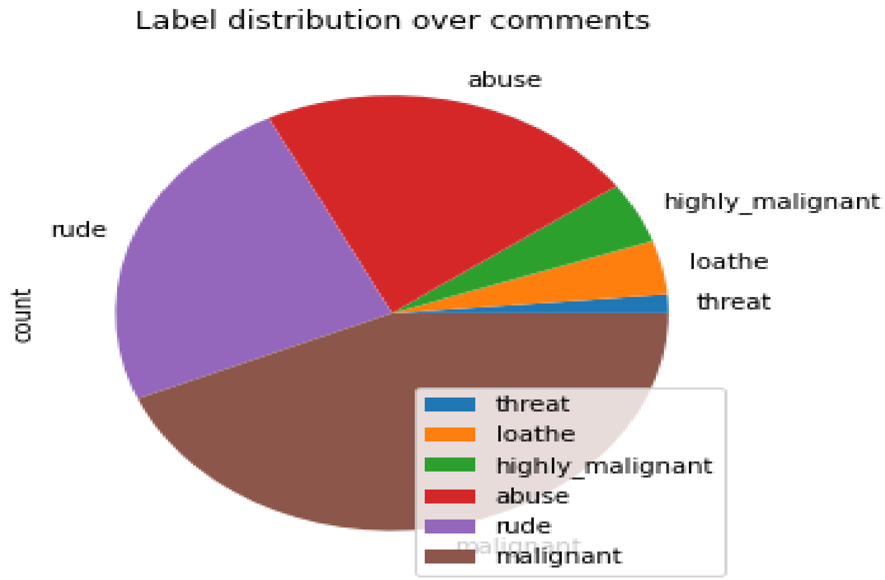


DATA PREPROCESSING STEPS

- Imported and applied wordnetlemmatizer and stopwords from Natural Language Toolkit in the dataset. In which Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. And stopwords are English words which don't add much meaning to a word.
- Created another variable called ‘Comment_length’.
- Plotted the count of maximum words from each category using wordcloud.



- Plotted a pie graph to understand the different categories.



Most of the messages are in the category of malignant, followed by rude and abuse.

- Converting the target variables to single variables as bad.
- Input variable is comment_text, inorder to convert the text to vector applied TF-IDF.

MODEL TRAINING AND EVALUATION

Trained the dataset on different models and below table explains the score:

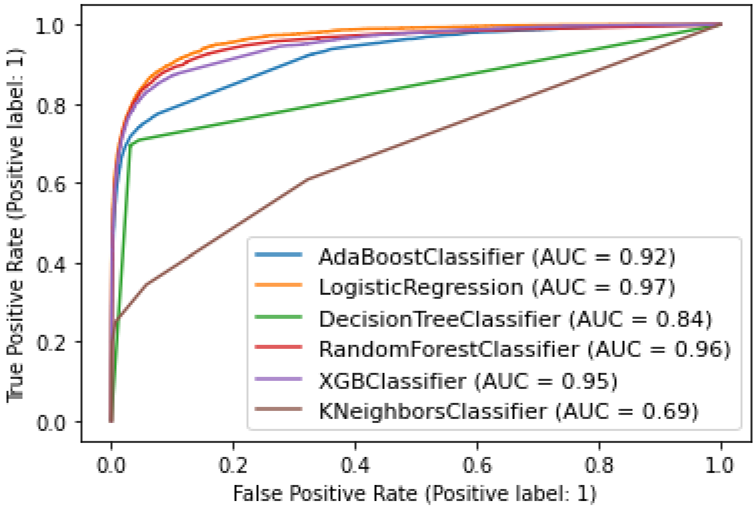
No.	Model	Accuracy	Cross-validation	Difference
1	Logistic Regression	95.50	95.64	0.14
2	Decision Tree	94.06	94.12	0.06
3	Random forest	95.56	95.65	0.09
4	XGBoost	95.23	95.36	0.13
5	Adaboost	94.56	94.59	0.03
6	KNN	91.67	91.35	0.32

From the above table, I have considered the Adaboost Classifier as the best model because the difference is less,, which means the possibility of overfitting is less.

After deciding the best model I have done Hyper Parameter Tuning by considering learning rate as 0.05 and n_estimators as 100..

After HPT the accuracy increased to **94.96%**.

ROC-AUC CURVE



Applied the same transformations on test.csv and predicted the result.

CONCLUSION

- Transformation techniques like lemmatization, TF-IDF method is applied to the dataset.
- Selected Adaboost Classifier as the best model.
- Adaboost classifier is performing with an accuracy of 94.96%