

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

RAHUL ADHWARIA

**ACKNOWLEDGMENT**

The internship opportunity I had with Flip Robo Technology was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this internship period.

I express my deepest thanks to Shubham Yadav, SME of internship for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make work easier. I choose this moment to acknowledge his contribution gratefully.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

**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 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.

* Motivation for the Problem Undertaken

From this project I get to learning many data science and ML topics. From this project I get to know how NLP and data science topic is worked in combination. In project enhance my hands on data science programing and theory parts.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In our project we have character data, therefore here we used classification. In this project I used Decision Tree Classifier. As it is not mentioned to try different models, therefore we use single algorithm because we are getting pretty good accuracy of 89% and it is enough to predict test data.

* Data Sources and their formats

The data is given by the company to analyse the reviews and comments, to work on its information, the client wants some predictions that could help them in further finding about comments by the reviews given in comment text column. In this we at the starting try to understand the features of dataset. In this we find the dimensions, type of data, etc.

* Data Preprocessing Done

For doing data preprocessing, we firstly import NLP library nltk, and then we split data into word tokenization, remove defined stopwords and also include some self-required stopword which we want to remove if present in data. After that we remove punctuations. And then we sent our split data into tf-idf method to convert words into numbers so that we can perform ML model building so that we can train and test our model to predict the best accurate solution if we pass any review through it.

* Data Inputs- Logic- Output Relationships

We split our dataset into independent and target variable, as we have to find the best model. We select Decision Tree Classifier. We train the model and then predict the test dataset output.

* State the set of assumptions (if any) related to the problem under consideration

As we are very much familiar to various algorithms, but maybe there is a very high chance that the best algorithm for the give dataset is something which is not used by me in this project. So in this project I assumed that the algorithm I have selected will give the best model but somehow there can be anyone who has used any other algorithm and getting accuracy more than my selected model (Decision Tree Classifier).

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

In the starting we load our train dataset. After that we did data cleaning, firstly we understand the feature of dataset, then we plot some required graph which is shown below and make all letter in lower case etc. Then we import NLP library, and we split sentence into word tokenization, then we remove stopwords and punctuations. Then we sent our data to tf-idf method to convert into numbers, so that we can build ML model which will use a best fit model in future, if I have comment text or reviews then I can predict comment text or reviews respectively. In our case we find Decision tree Classifier as the best model.

Then we load our test dataset and transform it by tf-idf method and predict the required data as per the train model building.

* Testing of Identified Approaches (Algorithms)

As it is not mentioned in problem statement that we have to try different algorithm. So we build only one model as Decision tree classifier which is giving pretty good accuracy of 89%, which is enough to predict the test dataset.

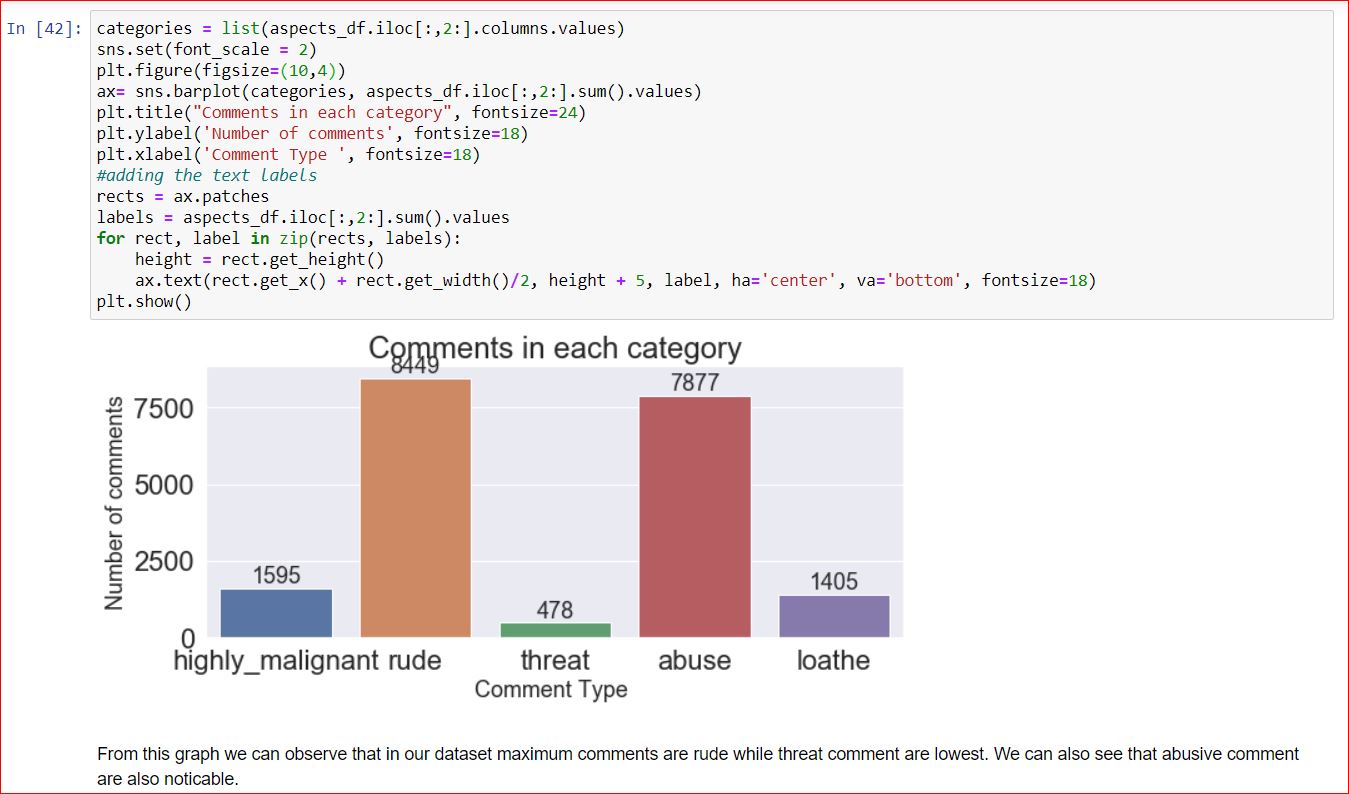
* Run and Evaluate selected models

As in our target variable we have categorical values, and we know that this type of dataset falls under classification. So we use Decision Tree Classifier to predict the test data. In this algorithm we are getting accuracy of 89%.

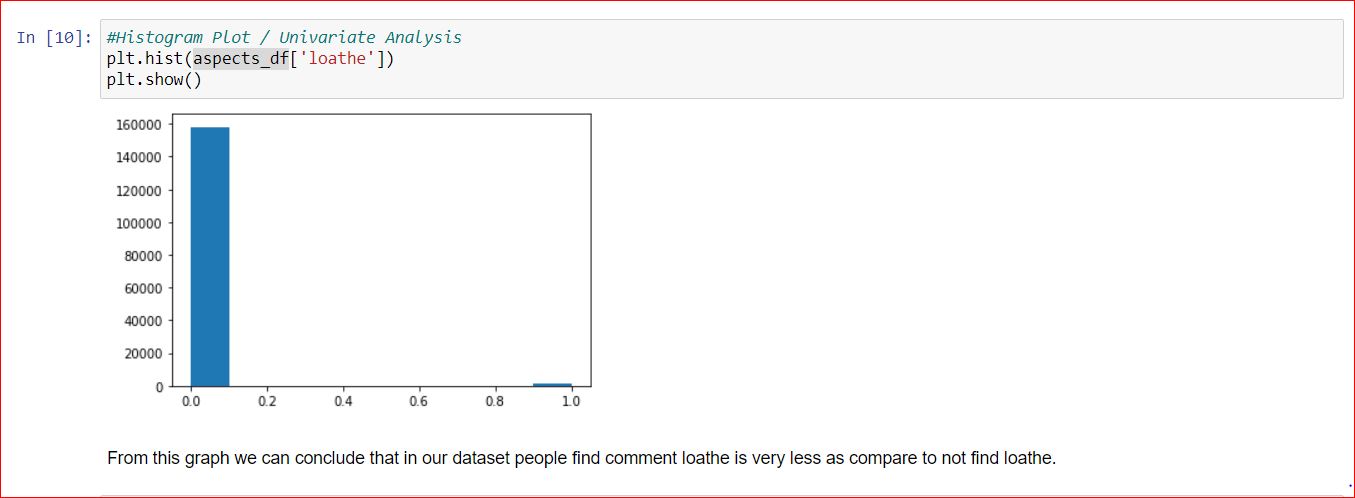


* Visualizations

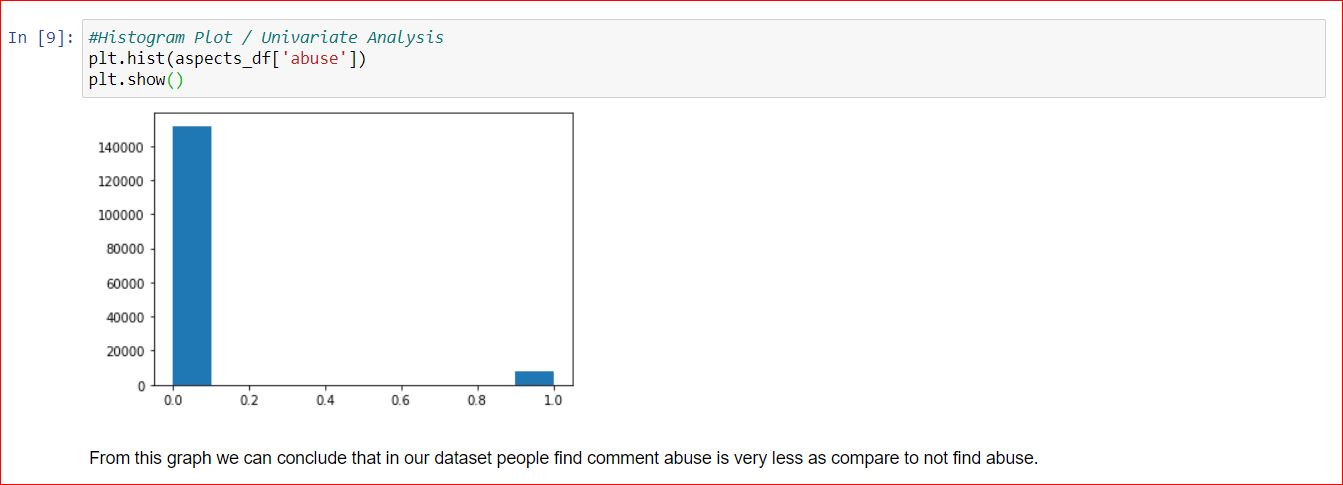
1. From this graph we can observe that in our dataset maximum comments are rude while threat comment are lowest. We can also see that abusive comment are also noticable.



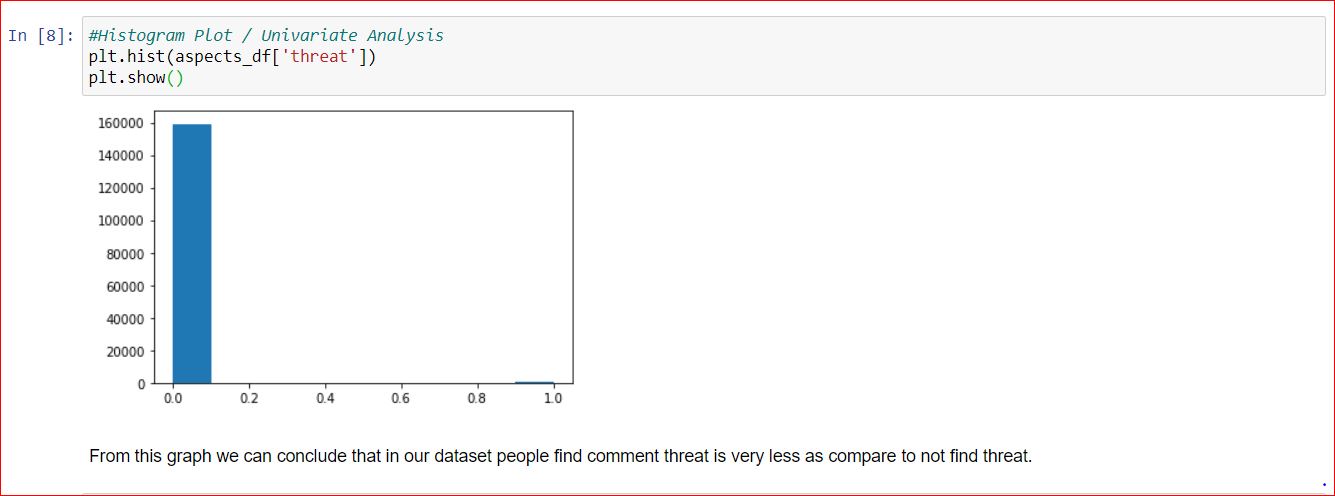
1. From this graph we can conclude that in our dataset people find comment loathe is very less as compare to not find loathe.



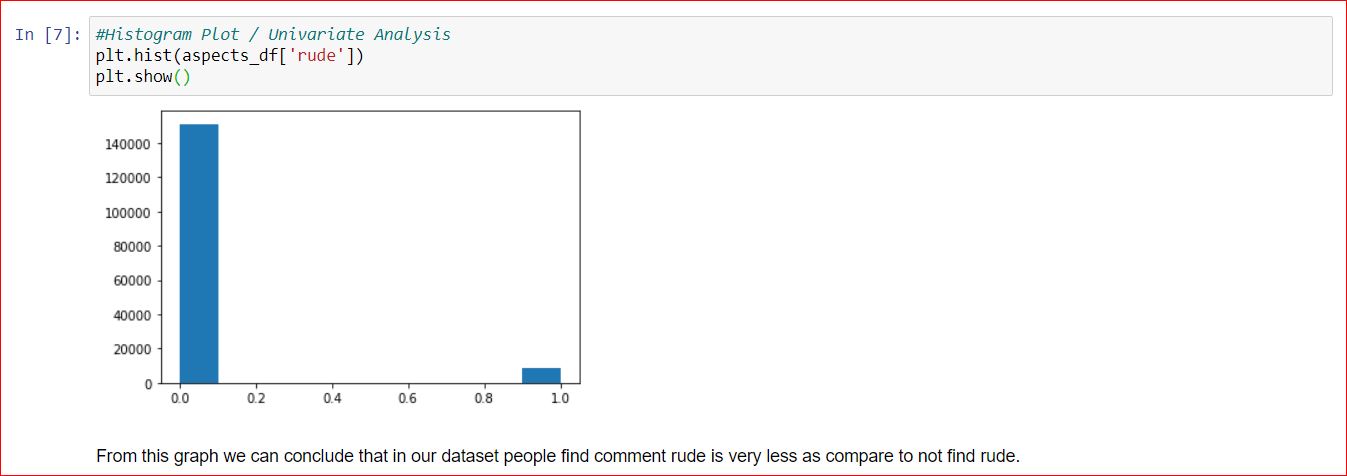
1. From this graph we can conclude that in our dataset people find comment abuse is very less as compare to not find abuse.



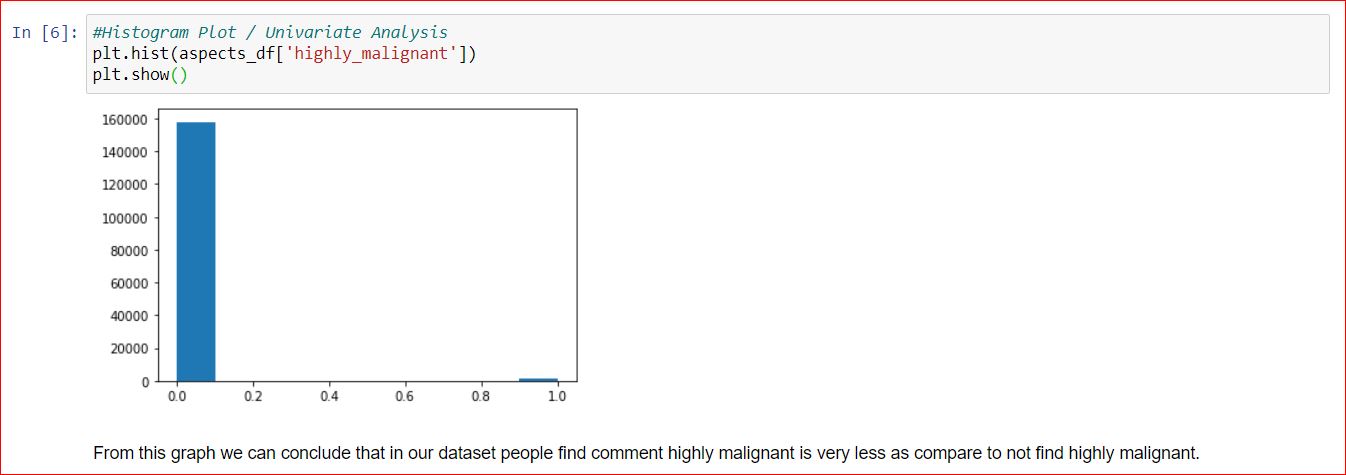
1. From this graph we can conclude that in our dataset people find comment threat is very less as compare to not find threat.



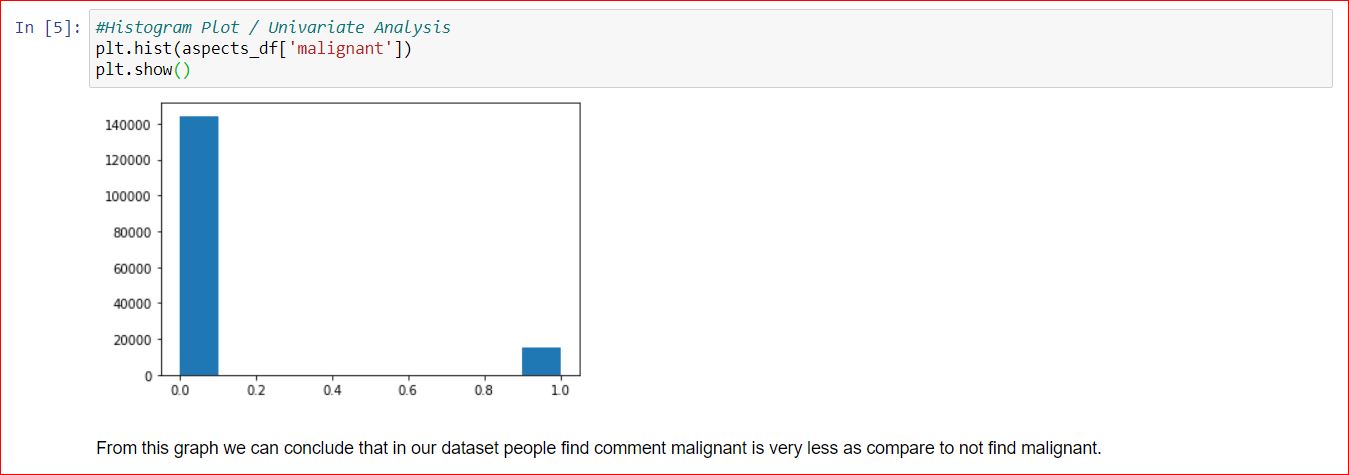
1. From this graph we can conclude that in our dataset people find comment rude is very less as compare to not find rude.



1. From this graph we can conclude that in our dataset people find comment highly malignant is very less as compare to not find highly malignant.



1. From this graph we can conclude that in our dataset people find comment malignant is very less as compare to not find malignant.



* Interpretation of the Results

From the exploratory Data analysis, we could generate insight on the data. How each of the features relates to the target. Also, it can be seen from the evaluation that Decision tree classifier with accuracy score of 89%.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this dataset we can conclude that as per our observation Decision Tree Classifier predict the target variable of our test dataset with a pretty good accuracy.

* Learning Outcomes of the Study in respect of Data Science

1. From graph 1 we can conclude that in our dataset people find comment malignant is very less as compare to not find malignant.
2. From graph 2 we can conclude that in our dataset people find comment highly malignant is very less as compare to not find highly malignant.
3. From graph 3 we can conclude that in our dataset people find comment rude is very less as compare to not find rude.
4. From graph 4 we can conclude that in our dataset people find comment threat is very less as compare to not find threat.
5. From graph 5 we can conclude that in our dataset people find comment abuse is very less as compare to not find abuse.
6. From graph 6 we can conclude that in our dataset people find comment loathes is very less as compare to not find loathe.
7. From graph 7 we can observe that in our dataset maximum comments are rude while threat comment is lowest. We can also see that abusive comment is also noticable.

* Limitations of this work and Scope for Future Work

This work is able to fulfil the future scope of our client as they want some predictions that could help them in further investigation and improvement in selection of ratings.