

MALIGNANT COMMENT CLASSIFICATION PROJECT

Submitted by:

SANJAY SIVAKUMAR

**ACKNOWLEDGMENT**

In this project I have done only the analysis which I learned by doing hands-on during my learning period of NLP. So, In this learning period I have mostly referred youtube videos, blogs etc related to NLP.

**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.

* Conceptual Background of the Domain Problem

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.

* 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 unoffensive, but “u are an idiot” is clearly offensive.

* Motivation for the Problem Undertaken

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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

The basic target for this problem is to classify the hate and abusive comments from the normal comments. Thus we use various NLP techniques for this process.

* Data Sources and their formats

The dataset consists of two columns one input and one target variable.



* Data Pre-processing Done

1.Firstly,the most important column,the comment\_text needs to be processed thoroughly .

2.Then, cleaning the review column as their will be many unwanted words such as special characters, stop words etc...

3.Finally converting all the current words to lower case for the purpose of model building.

* Data Inputs- Logic- Output Relationships

In this dataset we have 8 columns one is our input column which is used for the NLP process and the remaining is our target variables .

* Hardware and Software Requirements and Tools Used

For this project I have used python as programming language and JupyterNotebook as its tool. The libraries and packages I have used in this project are,



1.pandas- for data framing the dataset

2.numpy- for any analytical functions

3.seaborn and matplotlib-for visualization process

4.And for the NLP process we use TfidVectorizer to convert our review data from string to vector.

5.train\_test\_split- for splitting data into training and testing

6.And I have imported classification algorithms such as naïve bayes algorithms, RandomForest classifier etc… for model building.

7.And finally imported metrics for classification problem for proper model selection.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

1.Cleaning the review the input column as their will be many unwanted words such as special characters, stop words etc...

2.Finally converting all the current words to lower case for the purpose of model building.

3.Using TfidVectorizer which is the NLP process to convert the string data in review column to vector format for model building purpose.

* Testing of Identified Approaches (Algorithms)

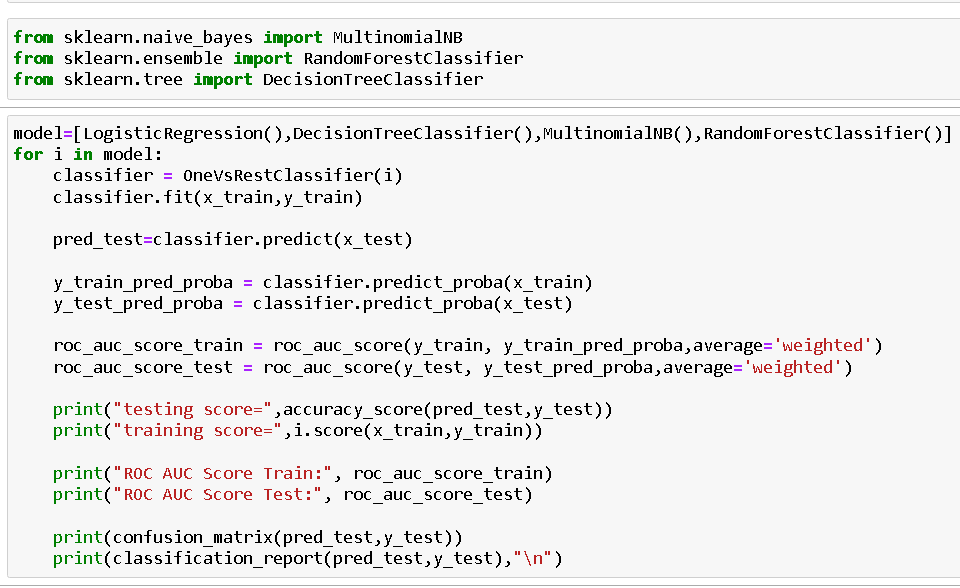
1.Naive Bayes Algorithms

2.Random Forest Classifier

3.Decision Tree Classifier

4.Logistic Regression

* Run and Evaluate selected models



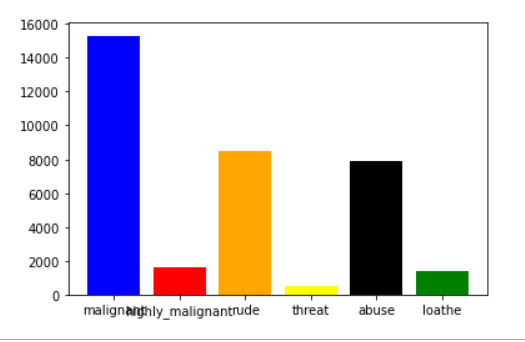
* Key Metrics for success in solving problem under consideration

1.Accuracy score- to know the accuracy of the model

2.Confusion matrix- provides us the number of correct and incorrect predictions

3.Classification report-provides us the precision and recall value of the model

* Visualizations



-Frequency of various types of abusive comments.



* Interpretation of the Results

The model’s AUC-ROC score is pretty good in these scenario. As this is my second NLP project I still think that there is many further techniques to learn in this project.

**CONCLUSION**

* Key Findings and Conclusions of the Study

As it is only my second project in NLP, I have found that there is many unwanted words or characters in the comment\_text columns as it all will make harder for the model building process and hence it all should be removed properly.

* Learning Outcomes of the Study in respect of Data Science

By this project I have learned that there is a sea of things to learn in this Data Science. And, thus I have learned the basics of NLP while doing this project and now I am very excited to further explore this field.

* Limitations of this work and Scope for Future Work

After my final model I got an accuracy which is pretty good in this scenario and this accuracy can be further increased by adding more by using advanced feature Engineering to this dataset.