

Malignant Comment Classifier

Submitted by:

Puspraj Kumar

**ACKNOWLEDGMENT**

I am going to mention that all references, first web site is ‘towardsdatascience.com’, ‘researchgate.net’ and some other website. Sources of data is amazon.com so, these are helped me and guided me in completion of the project.

**INTRODUCTION**

* Business Problem Framing

This problem is related to the real world in this sense, 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.

I tried this through this project, to build a Natural language Processing model that can classify the malignant comment or not. We should be able to read and understand which comment are malignant and which are not.

* Review of Literature

This is a comprehensive summary of the research done on the topic. Firstly, I have dropped those features which does not affect on the target feature. Through the Len () function, I have created one more feature for comment length which give the length of data available in these features. After doing this I have converted all the data in lower case and remove all regular expressions. I have also removed all stop words.

* Motivation for the Problem Undertaken

The objective behind to make this project is how we detect the which comment malignant and which are not. Rude comment is written on social media platform also a big problem now a days they affect some time very deeply. Because of this people goes through stress. So, I was wanted learn for these topics, through this project I have definitely learn something in this filed. So that is the motivation for the problem undertaken.

**Analytical Problem Framing**

* Data Sources and their formats

Source of data, their origins, and their formats. Data is given by Flip Robo Technology SME in the .zip file. Files have train file and test file which is in .CSV format. The train data file has 8 features.

• Data Pre-processing Done

Data pre-processing done here for doing this project, removed all null data, regular expression, stop words and punctuation.

• Data Inputs- Logic- Output Relationships

The relationship behind the data input, its format, the logic in between and the output. Data input is train and format are in csv form. The logic in between and the output is removing regular expression and stop words form the data set. Over on these thing input affects the output.

• Hardware and Software Requirements and Tools Used

The hardware and software requirements along with the tools, libraries and packages used. Hardware I have use for doing this project is window 10 operating system. Platform I have used for doing this project is jupyter Notebook. Libraries I have used is nltk and word cloud.

**Model/s Development and Evaluation**

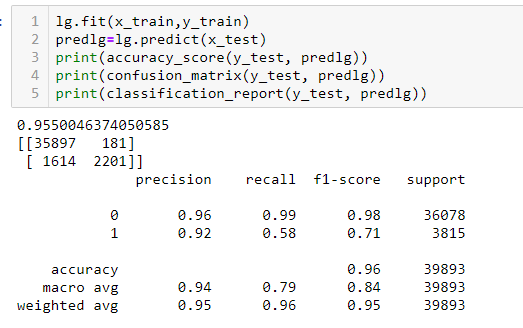
* Identification of possible problem-solving approaches (methods)

I am going to describe the approaches I have followed, for solving this problem. Firstly, I have dropped those features which not effects on the label feature. After doing this I have create one newer feature which have length of data in feature comment\_text. I have also removed all regular expression, stop words, punctuation or special character.

* Testing of Identified Approaches (Algorithms)

For the testing I have used ‘model. fit’ function in which I have tested on multiple models. Almost for every model I have got more than 85% of accuracy, precision and recall. Which is good.

* Run and Evaluate selected model



Snapshot of their code and what were the results observed over different evaluation metrics.

* Visualizations

The plots made along with their pictures and what were the inferences and observations obtained from those. For plotting I have used matplotlib, seaborn and Word Cloud library from these libraries I have showed before the pre-processing and after pre-processing on length of data feature. Through word cloud method I have showed the rude word and right words.

* Interpretation of the Results

Summary of what results were interpreted from the visualizations, pre-processing and modelling. From the visualization I have showed done above from them which is use full or not. I have done here pre-processing is remove regular expression, stop words and special character. Model I have used Multinomial NB, Decision tree, K-Neighbours and Random Forest.

**CONCLUSION**

* Key Findings and Conclusions of the Study

I am going to identifies key findings, inferences, observations from the whole problem. Removing regular expression and stop words is kye form this project.

* Learning Outcomes of the Study in respect of Data Science

Firstly, in NLP we learn how to deal with text or article, audio file data. I don’t think for doing this project I have faced difficult challenges. So, these things I have learned from this.

• Limitations of this work and Scope for Future Work

Limitation for doing this project is that when I have selected multiple feature fit transform in TfidfVectorizer it gives error. That’s why I have dropped the headline feature just above the vectorization.