

Malignant-Comments-Classifier Project

Submitted by:

MUKUL DHAMIJA

**ACKNOWLEDGMENT**

I would like to **express my special thanks of gratitude to** SME Kushboo garg who gave us the golden opportunity to do this wonderful project on the topic Malignant-Comments-Classifier Project. And I am also thankful to my mentor shankargouda tegginmani and data trained team who provide me useful material to solve this problem

**INTRODUCTION**

* Business Problem Framing

The background for the problem originates from the multitude of online forums, where-in people participate actively and make comments. As the comments some times may be abusive, insulting or even hate-based, it becomes the responsibility of the hosting organizations to ensure that these conversations are not of negative type.

* Conceptual Background of the Domain Problem

There are so many social networking sites such as twitter , facebook where huge crowd have freedom to comment by their own choice this models helps to classify either the comment is toxic or not

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

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 behavior.  
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 un offensive, 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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

First we have to do data cleaning by using EDA process with the help of numpy and panda various numerical value can be calculated usinf describe command we can calculate mean std min max ,25%,50%,75%.with the help of corr command we can calculated probability distribution

* Data Sources and their formats

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000samples. 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 0denotes 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 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.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

For model development we have to use train test split model library so that I can able to select algorithm to calculate accuracy value .

* Testing of Identified Approaches (Algorithms)

Various models are

1. Logistic regression
2. Decision tree classifier
3. Random forest classifier
4. Xgb boost classifier
5. Random forest classifier
6. Knn neighbour classifier

* Run and Evaluate selected models

With the the help split test train library we can calculate accuracy matrix an we can run it

* Key Metrics for success in solving problem under consideration

.

After getting matrix value we can compare the values more the values then we will select that model

**CONCLUSION**

* Key Findings and Conclusions of the Study

Comparison between confusion matrix more the value closer to one then model has to be selected

* Learning Outcomes of the Study in respect of Data Science

We can calculate which comment is toxic which one is not toxic and avoid cyber bullying