Icon

Description automatically generated

Malignant Comment Classifier

Submitted by:

Sarita shukla

**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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We used histogram and count plot to see the data distribution in each column, we also used subplots and bar graphs for comparing the data present in different columns. We also used heatmap to plot correlation between target variables and individual bar graphs for representing them in a more detailed manner.

* Data Sources and their formats

The data is split into two parts, train and test. The train set has 159571 rows and 8 columns, while the test set has 153164 rows with 2 columns. Hence, there were 6 target variables, i.e., 6 different categories, these were, malignant, highly\_malignant, rude, threat, abuse, loathe.

* Data Preprocessing Done

There were no null or missing values present in the data sets. The id column didn’t provide us with any information regarding the manner of the tweets, hence it was dropped. Since the only feature available to us was in the format of text data, necessary nlp processes were used to clean the data and make it ready for model building. These tests included lower casing the texts, removal of punctuations and stopwords, lemmatising the word to its root form and finally converting them into vectors for model processing.

* Data Inputs- Logic- Output Relationships

The data was built on the relation between the contents of the tweets and the nature of the tweets, and how the words used in tweets enabled us to understand the intentions of the writer.

* Hardware and Software Requirements and Tools Used

All of the work in this project was done on Jupyter notebook. We used pandas and NumPy for working on data and using all the basic mathematical functions on it. We also used matplotlib.pyplot and seaborn libraries for data visualization, also wordcloud library for making word clouds. We used sci-kit learn and nltk corpus library for data preprocessing and importing different models like Logistic Regression, MultinomialNB, Decision Tree Classifier, Random Forest Classifier.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

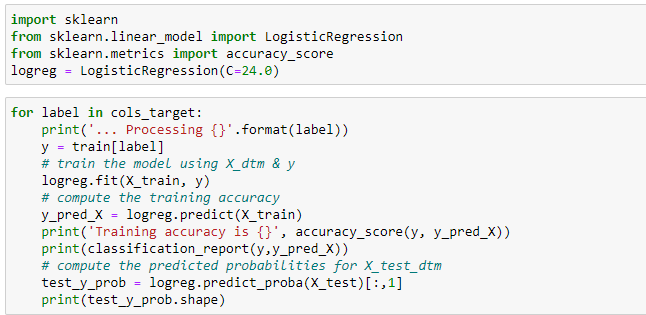
For converting words into vectors TfidfVectorizer was used from sklearn feature extraction library. Since we had 6 different target variables we used a for loop to loop over the dataset and give predictions for individual target variable.

* Testing of Identified Approaches (Algorithms)

We used Logistic Regression, MultinomialNB, Decision Tree Classifier and Random Forest Classifier.

* Run and Evaluate selected models

We used Logistic Regression, MultinomialNB, Decision Tree Classifier and Random Forest Classifier, below is the snapshot of code that we used to implement our model.

A screenshot of a computer

Description automatically generated with low confidenceTable, calendar

Description automatically generated

Text, letter

Description automatically generatedA picture containing table

Description automatically generatedTable

Description automatically generated with medium confidence

A picture containing calendar

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

A picture containing text

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

* Key Metrics for success in solving problem under consideration

For evaluation of our models, we used accuracy score, precision, recall and f1 score. We presented these scores with the help of classification report

* Visualizations

We used subplots to study the data distribution, heatmap to study correlation between different target variables, bar graphs to individually study the correlation in each target variable and word cloud to study the most prominent words in each category.

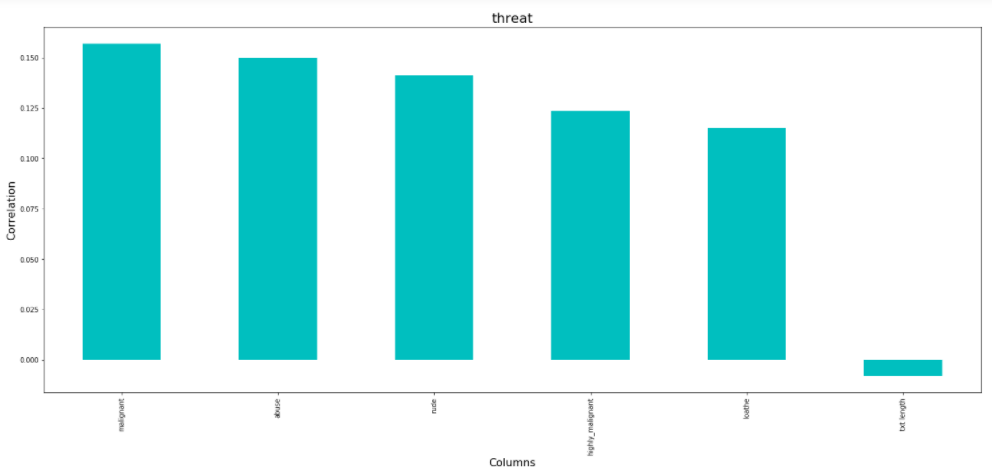
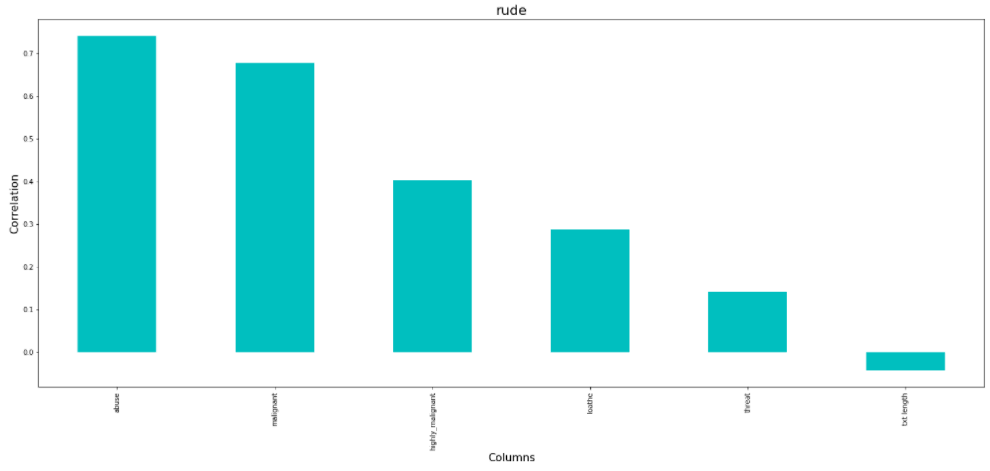


Chart

Description automatically generated

Graphical user interface, chart, bar chart

Description automatically generated with medium confidenceChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated

* Interpretation of the Results

We observed from the subplots how unbalanced our datasets are, from the heatmaps and individual graphs we can notice that rude and abuse have the highest correlation around 0.74, while the threat column had least correlation with other columns with values ranging between 0.12 - 0.16.

After preprocessing the data and building our model on the dataset, our accuracy scores were Logistic Regression - 0.964, 0.9924, 0.9841, 0.9984, 0.9761, 0.9945 for malignant, highly\_malignant, rude, threat, abuse, loathe respectively, MultinomialNB- 0.9519, 0.9908, 0.9749, 0.9970, 0.9700, 0.9918 for malignant, highly\_malignant, rude, threat, abuse, loathe respectively, Decision Tree Classifier - 0.9974, 0.9989, 0.9985, 0.9998, 0.9977, 0.9994 for malignant, highly\_malignant, rude, threat, abuse, loathe respectively, and Random Forest Classifier - 0.9974, 0.9989, 0.9985, 0.9998, 0.9977, 0.9994 for malignant, highly\_malignant, rude, threat, abuse, loathe respectively, our decision tree model and random forest model had same accuracy upto 4 decimal digits but decision tree model had slightly better precision and recalls, hence we decided to finalize our Decision Tree model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project, we studied how people behave on different social media platforms where they had no jurisdiction to govern their antics .

Using predictive modeling and EDA we can determine which comments can be labelled as toxic or malignant and the severity of the comment by just processing the contents of the content.

After all the data processing and data analysis, we manage to build a model which will give us fairly accurate results.

* Learning Outcomes of the Study in respect of Data Science

While working on this project I learned about natural language processing and how it can be so useful for different social media platforms like twitter, instagram, etc.

* Limitations of this work and Scope for Future Work

Even though the accuracy and other scores were pretty satisfactory, this was partly due to the imbalanced nature of the dataset, but still it gave us fairly satisfying results.