

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

Mayank Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to take a step in this project. I gone through google and Kaggle for the references and my old projects helps me understanding the way to solve this use case.

**INTRODUCTION**

* Business Problem Framing
  1. 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.
  2. 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.
  3. In social media also there are various comments which are offensive and need to ban so this model can able to detect the bad words and help cyber department to block or report.
* Conceptual Background of the Domain Problem

1. There are various aspects and point of view to understand the domain of this type of problem as there are many social handles of normal and celebrity also which face both type of comments.
2. So, primarily we must focus which social handles is used mostly among all and what are the moto of particular social handles who is facing this type of comments.
3. There are lot of social issues happen in the world which get trend and bad comments spread so we must also focus on trend news or current popular activity.

* Review of Literature
  1. 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 must 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.
  2. Recently there are lot of bad and tags trend on SSR case and million of mixed type of comments are in trend, so there are many cases in which comments play a crucial role to gain and loss of trend
* Motivation for the Problem Undertaken

Because of the most demanding and most exploratory technology i.e. NLP and specifically using machine learning in this type of problem can help me to learn more and explore me more. The following problem statement is also a useful use case because as said earlier comments classification is becoming more important now a days.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

1. Project contain train and test dataset.
2. In train data set there are 159,571 rows and 8 columns.
3. In test data set it is like 153,164 rows and 2 columns.
4. There are no null values in the dataset
5. Most of the data are numeric in nature which are binary.
6. Comments is object in nature and consist of text.
7. Overall memory usage for train and test is around 15MB.

* Data Sources and their formats

Train Data

Text, table

Description automatically generated

Test Data

Text

Description automatically generated

* Data Pre-processing Done

1. As there are no null values in the dataset, but as the comment column in text format so there require lot of text pre-processing.
2. Percentage of unlabelled comments or good comments is 89.832.
3. Percentage of all bad comments 0.0194.
4. As, number of offensive comments are in very less in number.

* Data Inputs- Logic- Output Relationships

1. There are features which are highly correlated
2. It is possible because one comment can be classified into multiple categories
3. One comment can be rude and abuse at same time.
4. Most of the text length are within 500 characters, with some up to 5,000 characters long

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

For the easiness of the training and predicting I assumed and make a label of bad and good words.

Means all the bad comments are classify as bad word and all good comments whose value are 0 are classified as good comments

This will help the model to classify the comment easily also I had made a separate model which helps to predict the prob of comment text that in which column it has high prob.

* Hardware and Software Requirements and Tools Used

Graphical user interface, text, application

Description automatically generated

Above are the libraries which I used to pre-process, predict and visualize the project.

Pandas - It is used to play with data frame and helps in to get more insight of the data, like describing the data and the types of the all features.

Seaborn, Matplotlib – This is a visualization library which helps to plot different type of bar charts and line charts which help to get more information from the dataset.

Sklearn – This library is most important library which include all the metrics, accuracy, and the model which help to predict the output all classification and regression algorithms followed by all vectorizer which is responsible for creation of bag of words.

Nltk – This is the library mainly responsible for performing NLP related tasks like stemming, text pre-processing etc.

Re – It is used for regular expression to find or split the text.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I followed the path of lifecycle of Data Science which includes:

Diagram

Description automatically generated

1. Read the use case and search from the google to know more about the use case and in which field or domain it will applicable more.
2. As data set is already there, so this step is excluded.
3. Data Preparation includes the data cleaning and describing the data which I followed text pre-processing.
4. EDA involves the visualization which is helpful to get more insight from the data and get to know about the mostly use comment and check the most frequent words.
5. Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and apply hyperparameter tuning.
6. Model Evaluation, for this I used confusion metrices and mainly focus on False negative and tried to reduce the False negative which is type 2 error and on F1 score as dataset is imbalance.

* Testing of Identified Approaches (Algorithms)

Decision Tree:

Table

Description automatically generated

Random Forest:

Table

Description automatically generated

XG Boost:

Table

Description automatically generated

* Run and Evaluate selected models
  1. In classification problem there are various metrics that are accuracy score, confusion matrix, classification repot, Roc Auc curve which help to check the efficiency of the model
  2. Which metrices is useful? Is also depend and vary on domain, so as per the use case we must predict that whether the customer is defaulter or not
  3. So, in this case accuracy score is good but most important is confusion matrix in which we must decrease the False Positive that is type 2 error.
* Visualizations

Mostly I used seaborn, word cloud and matplotlib for visualization and EDA

Chart, pie chart

Description automatically generated

Most of the comment are of malignant and rude.

Chart, bar chart

Description automatically generated

Count of comment in each comment.

Chart

Description automatically generated

Abuse and malignant, rude and abuse are highly correlated.

Chart, histogram

Description automatically generated

Count the number of texts in each comment, mostly fall under 1000.

Text

Description automatically generated

These are some most good comments from dataset.

**CONCLUSION**

* Key Findings and Conclusions of the Study

1. In this project there are some variables like malignant and rude which are highly correlated it is possible because one comment text may have combination of multiple features.
2. Removing the column id does not impact the model training.
3. Using Tree, model can reduce the false negative values

Table

Description automatically generated

These are some of words which are classified by Logistic Regression which make comment as bad label.

* Learning Outcomes of the Study in respect of Data Science

1. While implementing this project most of the time is taken doing the text pre-processing
2. Gain knowledge and get more insight of various stemmer and vectorizer.
3. Random forest is well suitable for this project as it used tree internally and it used multiple weak learner and generate the strong model and generate low bias and low variance model.
4. Data visualization is also I learnt from this project various plot and graph I plot which help to get more insight of data.

* Limitations of this work and Scope for Future Work
* It has future scope in various use cases likewise in election, social media etc, where every day there are multi offensive comments spread.
* So, in the future it may use very well to easily classify the comments as bad or good.