Icon

Description automatically generated

MALIGNANT COMMENTS CLASSIFIER PROJECT.

Submitted by:

ABHISHEK GUPTA

**ACKNOWLEDGMENT**

I would like to acknowledge everyone who played a role in my academic Accomplishments. First of all, my guided teacher and employer who helped me in doing in this project. I have used my proper knowledge and guidance to work in this project at maximum capacity.

In this project I have include the research papers, various machine learning algorithms, websites which helped me to analyse the data and then take corrective measures in it and at last the given dataset from which I pulled out the relevant information.

INTRODUCTION

**Business Problem:**

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.

**Conceptual Background of the Domain Problem:**

The main domain and conceptual problem of the given data that is provided to us is indicating various definitions of the data description where the concluding column or the predictive column will be analysed by applying various machine learning algorithms such as Logistic Regression, K Neighbours Classifier, Decision Tree Classifier, Gradient Boosting Classifier etc.

These data columns are the indications of a particular reason which helps the client to make a future prediction based on the data analysis of past records. The size of the train data is 159571 rows × 8 columns and the size of test data is 153164 rows × 2 columns which describes the data records which are as follows.

Most of the data contains text format values except 6 columns that is our label columns which indicates the results of offensive category and non-offensive category based upon the comment-text column which is the sole indicator for the predictive analysis of our given dataset.

The given dataset does not contain any null values for any column as seen in the heatmap.

This data description is useful for the better understanding of my project which will be calculated after analyzing the different columns and understanding their co-relation with the label column and overall efficiency of the given dataset.

**Review of Literature**

This is a comprehensive summary of my research done which is based on the various factors such as data cleaning, doing exploratory data analysis, finding out the null values and then correcting them with the next best alternative, detecting outliers, pre-processing the data, applying various machine learning algorithms and then finally deciding, implementing that algorithms which is giving the maximum accuracy with the least number of errors.

In the process of analysing the data I have used various methods and coding techniques to get the better understanding of the data.

**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.

This analysis is based upon the past records for the dataset given and no. of factors to analyse each comment as offensive or non-offensive in nature to a particular user so that in future we can take corrective measure to minimize the hatred and cyberbullying.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

The mathematical analysis of the given data is understandable in the following way:

* There are too much of text data present in the data which needs to be classify and cleaning from all the errors such as invariant signs, unnecessary punctuation, emails id, mobile numbers etc.
* There is no null value present in our dataset as we can see in the heatmap for the same.
* Data includes 6 label columns with the binary values as “1” being the offensive comment and “0” as a non- offensive comment.

**Data Sources and their formats:**

In the given data set the data description defines the basis of each column that ultimately affect the effectiveness and efficiencies of the label column as it indicates the success or the failure in terms of spreading hatred and do cyberbullying in the society and identifying the comments as to be malignant or non-malignant.

The data description and their formats is as follows which helps us to get a better understanding of the problem statement in analysing the given data.:

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **id** | A unique id aligned with each comment text. |
| **comment\_text** | It includes the comment text. |
| **malignant** | It is a column with binary values depicting which comments are malignant in nature. |
| **highly\_malignant** | Binary column with labels for highly malignant text. |
| **rude** | Binary column with labels for comments that are rude in nature. |
| **threat** | Binary column with labels for threatening context in the comments. |
| **abuse** | Binary column with labels with abusive behaviour. |
| **loathe** | Label to comments that are full of loathe and hatred. |

The given columns describes the meaning of the data that belongs to the user of varous social media platform.

The information regarding the data types , values counts and the null values are as follows:

|  |
| --- |
| # Column Non-Null Count Dtype  --- ------ -------------- -----  0 id 159571 non-null object  1 comment\_text 159571 non-null object  2 malignant 159571 non-null int64  3 highly\_malignant 159571 non-null int64  4 rude 159571 non-null int64  5 threat 159571 non-null int64  6 abuse 159571 non-null int64  7 loathe 159571 non-null int64 |

**Data Preprocessing Done:**

In this section I have covered a part which includes the data cleaning, assumption made, modification and updating of the data to get the relevant insights or the information out of it. This process is necessary to get the more useful data which can be used, removes the unnecessary data which hinders the process of data visualizations and performing the model selection part where we can decide the future credit amount defaults based upon the history. The process is divided into various steps which are performed to put our data in normal distribution at possible.

The steps are as follows:

1. After checking the comments\_text column it indicates that all the values are of the nature of simple text which needs to be clean to draw a meaning conclusion out of the data visualization for all the given dataset.so in this to clean the data as per our understanding of the data we can perform various techniques to get the clean length of this particular column.
2. To get the clean length for the comment\_text column I have remove the various unnecessary elements out of the data column such as email id’s, web addresses, various money symbols, mobile number etc to define the column in more presentable format for the better understanding of our given model.
3. After the removal of all these elements I have remove various punctuation attached to the comment\_text column , removed leading and trailing whitespaces , adjust the spaces in more define way. By applying the thses method, we can distinguish between the comment\_text column , and the clean length column for that data. After this it is easy to understand and visualize the given dataset.
4. After this I have used the technique of stopwords.words to move the unnecessary English words which can hinder the correct definition of our dataset to get the data in somewhat normal distribution.

**Data Inputs- Logic- Output Relationships:**

In this section I have conclude that most of the data column are directly co-related to its output variable that is label column which helps in deciding the offensive or non-offensive nature of comment.

If I look into the displot figure between the malignant comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated with medium confidence

This is the relationship between the malignant and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between malignant comment and clean length of comments

If I look into the displot figure between the highly\_malignant comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between highly\_malignant and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between highly\_malignant and clean length of comments.

If I look into the displot figure between the rude comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between Rude and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between Rude and clean length of comments.

If I look into the displot figure between the Threat comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between Threat and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between Threat and clean length of comments.

If I look into the displot figure between the Abuse comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between Abuse and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between Abuse and clean length of comments.

If I look into the displot figure between the Loathe comments as per the clean and unclean length of comment\_text column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the comment\_text column and malignant column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between Loathe and unclean length of comments.

Graphical user interface, text

Description automatically generated

Relationship between loathe and clean length of comments.

**Hardware and Software Requirements and Tools Used:**

As for the hardware part the below details are as follows:

**Processor-** Intel(R) Core (TM) i5-6300 CPU @2.40GHz

**Installed Memory (RAM)-** 8.00 GB (7.82 Usable)

**System type-** 64-bit operating system, x64

**Libraries and packages used in python**:

1. import pandas as pd
2. import numpy as np
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from scipy.stats import zscore
6. import warnings
7. warnings.filterwarnings('ignore')
8. import string
9. import nltk
10. from nltk.corpus import stopwords
11. nltk.download('stopwords')
12. import os
13. get\_ipython().system('pip install wordcloud')
14. from wordcloud import WordCloud
15. from sklearn.feature\_extraction.text import TfidfVectorizer
16. from sklearn.naive\_bayes import MultinomialNB
17. from sklearn.linear\_model import LogisticRegression
18. from sklearn.svm import SVC
19. from sklearn.tree import DecisionTreeClassifier
20. from sklearn.neighbors import KNeighborsClassifier
21. from sklearn.model\_selection import train\_test\_split,cross\_val\_score
22. from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report,roc\_curve,roc\_auc\_score,auc
23. from sklearn.ensemble import GradientBoostingClassifier

All these tools and libraries are used to perform the model selection process and identifying which model is best fitted to my project with giving the least errors.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods):**

In this section it includes the approach that I have used to make statistical and analytical data analysis to solve the problem statement. Firstly, I have cleaned the data, finding out the statistical observations which is based on the given dataset.

Secondly doing an exploratory data analysis which describes the behaviour of each data columns with respect to its label target column and then finding out, the relationship between the input and the label target variable in the dataset which concludes the correlation of each columns to its label target.

Lastly, building the appropriate model by using different model building techniques to make the data out of extremities and then performing the machine learning algorithms to finding out which model is giving us the maximum accuracy with least errors in the confusion matrix as a performance metrics.

**Testing of Identified Approaches (Algorithms):**

As per this project where data is crucial to us and data size is too much, I have applied four basic algorithms as it is taking so much time to process it in my computer. The list of algorithms as follows:

* Logistics Regression
* Decision Tree Classifier
* Multinomial NB
* Gradient Boosting Classifier
* KNeighbours Classifier

**Evaluation of Selected Models:**

In the evaluation phase I have concluded that our model is performing and giving the best results in Logistics Regression with the least number of errors.

In other two classifier that is Decision tree and Gradient Bossting Classifier, I have observed that the model is getting the maximum accuracy score of more than 90% which is good but in KNN we are getting the lowest accuracy score of 88.97% with maximum number of errors whereas in our model for Logistics Regression we have attained the accuracy score of more than 95.42% with the least counts of errors as per the confusion matrix and classification report for the given dataset.

**Key Metrics for success in solving problem under consideration:**

As per the key performance metrics I have used accuracy score, confusion matrix, classification report, to determine the best fitted model as per the given dataset.

Key observations includes the working of best model with the algorithm of Logistics regression which given the maximum accuracy with the given accuracy score, confusion matrix and classification report.

**Visualizations:**

The plots and figures are as follows for the given dataset:

**Offensive comments as per the label columns**:

Text

Description automatically generated

Word cloud for malignant comments as offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for highly\_malignant comments as offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for rude comments as offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for threat comments as offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for abusive comments as offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for loathe comments as offensive comments as per comment\_text.

**Data visualization for non-offensive comments as per label columns.**

Text

Description automatically generated

Word cloud for malignant comments as non-offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for highly\_malignant comments as non-offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for rude comments as non-offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for threat comments as non-offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for abusive comments as non-offensive comments as per comment\_text.

Text

Description automatically generated

Word cloud for loathe comments as non-offensive comments as per comment\_text.

A picture containing shape

Description automatically generated

0 143346 non-offensive comments

1 16225 offensive comments

This indicates that exactly 16225 comments are offensive in nature whereas 143346 comments are non-offensive in nature. This output is after merging all the label columns and their values.

**CONCLUSION**

**Key Findings and Conclusions of the Study:**

In the given project there is a lot to know as the project was given to find the actual spreading of hatred and causes of cyberbullying in the society. 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.

While working in the project I come to know that most of the data is in text form and I have to clean the data first to gather the meaning information out of it and then perform various data visualization techniques to describe the data into more presentable format to the user or the client of this database.

The whole states the offensive and non-offensive comments as per different labels or measures of the society which will be predicted using different machine learning algorithms so that in future there is chance of implementing different techniques to manage the spread of hatred and take controllable measures for cyberbullying.