FLIP ROBO

Malignant Comments Classifier

Submitted By:

Shivanchal Asthana

ACKNOWLEDGMENT

It is my sensual gratification to present this report. Working on this project was an incredible experience that will have a tremendous impact on my career. I would like to express my sincere thanks to the company Flip Robo Technologies for a regular follow up and valuable suggestions provided throughout. They always been an origin of spark and direction. I also thank all the respondents who have given their valuable time, views and valid information for this project.

Shivanchal Asthana

INTRODUCTION

• Business Problem Framing:

We are required to model the comments classification malignant/highly-malignant with the available independent variables. This model will then be used by the management to understand to classify the comments based on input.

• Conceptual Background of the Domain 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.

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.

- ## Review of Literature:

  Now a days balancing an environment on social media platform is extremely important. Describing as “highly malignant and harmful passing through the medium of electronic text”, cyber bullying puts targets under attack from a barrage of degrading, threatening, and/or sexually explicit messages and images conveyed using web sites, instant messaging, blogs, chat rooms, cell phones, web sites, e-mail, and personal online profiles. Thus, the task of finding and removing toxic communication from social media forums is very crucial.

- ## ·Motivation for the Problem Undertaken:

This project helps me understand the toxic comments classification problem in social media platform, its customer comments. With the right set of datasets in hand I have built a model that helps the enterprise take the right decision that is whether to focus on a malignant/highly-malignant set of customers. This also motivate learn about text classification problem in social media platform in details. This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes many categories of comments, we can do good amount of data exploration and derive some interesting features using the comments text column available.

We built a model that can differentiate between comments and its categories.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem:

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.

. I have used a Random forest classifier model to classify the comments in terms malignant/highly malignant and also used cross validation to remove overfitting problem while predicted the correct outcome and validate the model.

Data sources are provided internally by the enterprise.

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

• Data Pre-processing:

In the data pre-processing stage, I have found out if there is any missing data in dataset, for a particular column if there are any outliers present and

how to handle the outliers. I have also found the total shape of the data set. I have also found out the dataset description using describe method. So, in this pre-processing process I have mainly cleansed the data and prepared the right set of data for further processing & for predicting the model.

• Data Inputs- Logic- Output Relationships:

To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. From this describe function that also known as Five-point summary analysis if there are any outliers are present for a particular column. Also five point summary analysis was done for the target variable to explore & understand the data in a better way.

- ### State the set of assumptions (if any) related to the problem under consideration:

Since all the dataset provided and defined properly so in this dataset, I assume malignant/highly malignant as the target variable for this project. Rest of the parameters are used as input variables.

- ### Hardware and Software Requirements and Tools Used:

For this particular dataset the Hardware is used Windows as operating system, and the software used are mainly Jupyter notebook for model building and various internal packages that are defined in the anaconda/jupyter notebook.

Model/s Development and Evaluation

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

For this particular project I have used different classification models to predict the outcome of this dataset. After the model implementation Random forest classifier method predicted the best outcome out of all the models in terms of accuracy score and also I have used cross validation to flag the problem related overfitting or selection bias for the dataset and hence we can use this model for further evaluation.

- Testing of Identified Approaches (Algorithms):

I have used mainly different classification methods to get the outcome of the house price prediction and 80% data used for training purpose and rest 20% are used for testing the prediction of the accuracy score for this machine learning model building process.

- Run and Evaluate selected models:

To predict the result of this dataset below are machine learning models used for evaluations.

Out of all the machine learning models used I have selected Random forest classifier model for further evaluation of this project.

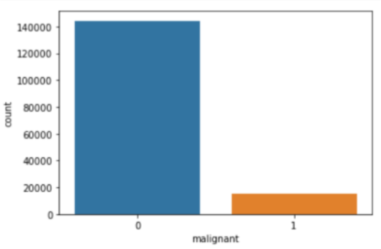
- Key Metrics for success in solving problem under consideration

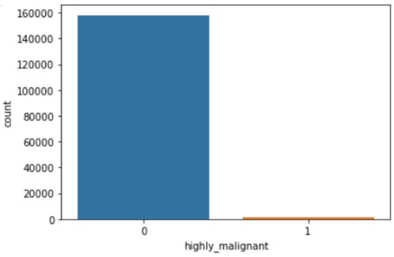
The key metrics that were mainly taken into consideration were the followings:

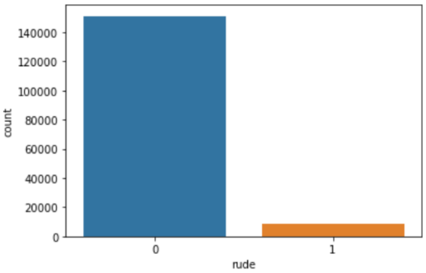
- Comments_text
- malignant
- highly_malignant
- rude
- threat
- abuse
- loathe

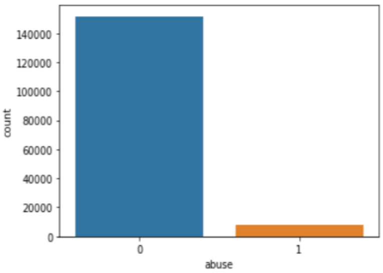
These are all the prime metrics under consideration.

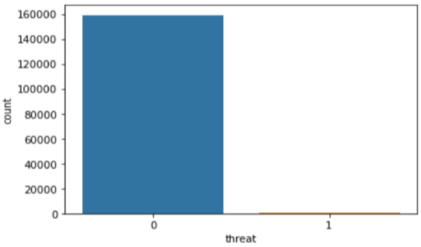
Visualizations:

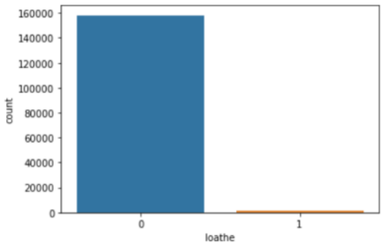


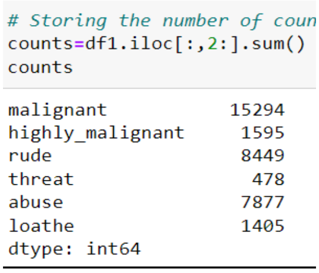






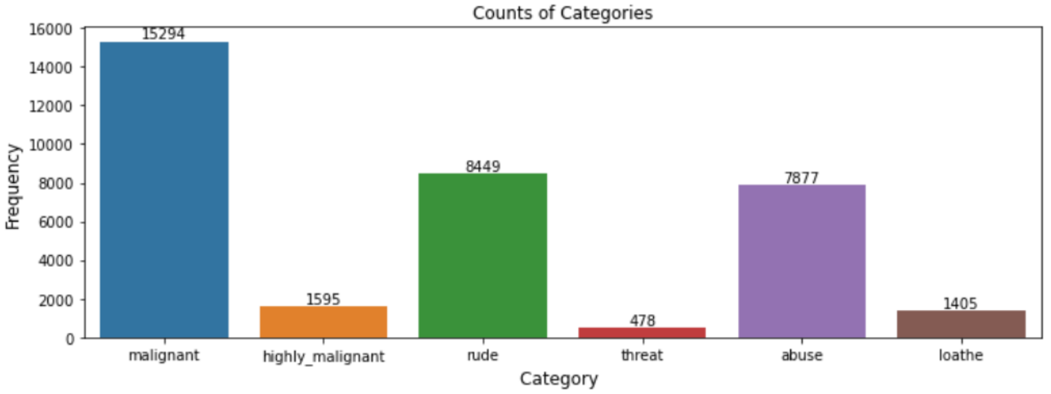


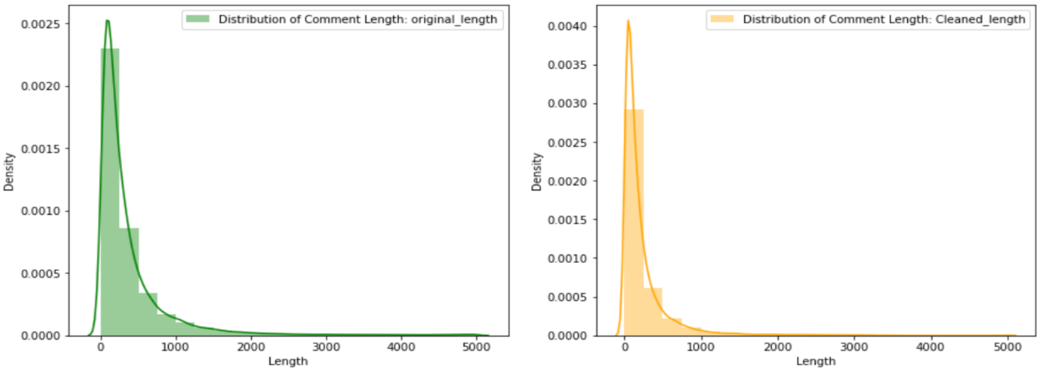


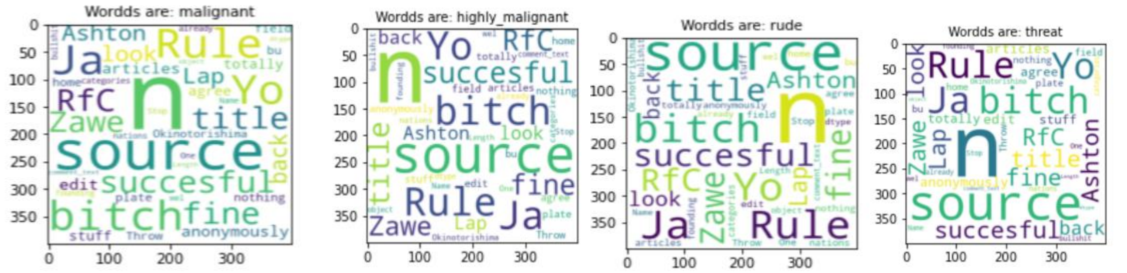


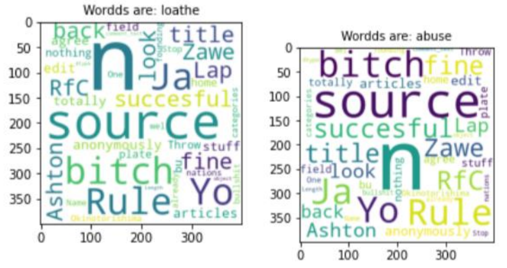
```
# Storing the number of coun
counts=df1.iloc[:,2:].sum()
counts
```

```
malignant           15294
highly_malignant     1595
rude                 8449
threat                478
abuse                7877
loathe               1405
dtype: int64
```

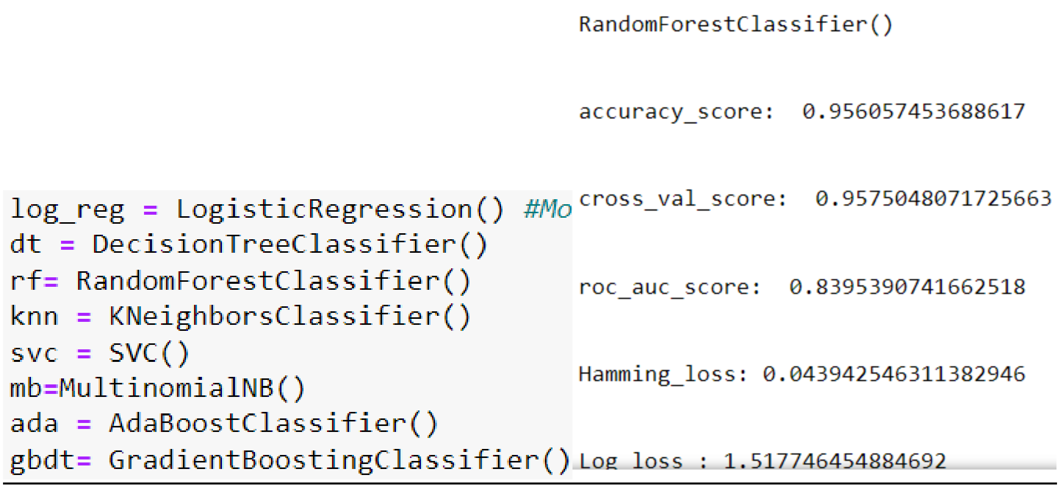






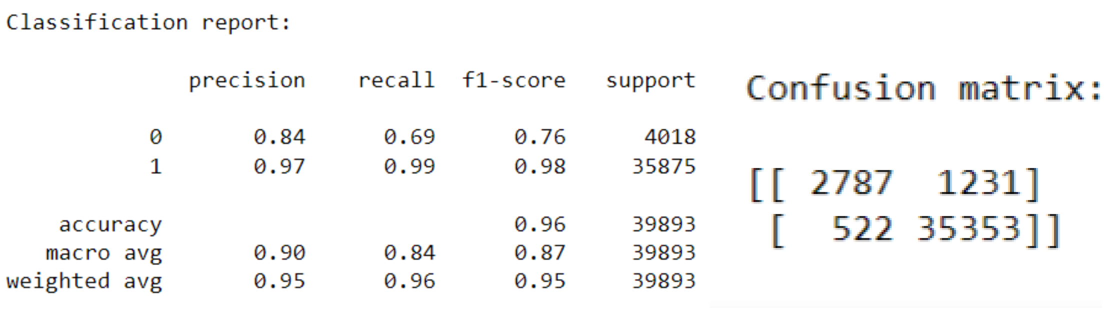


1. We can see here, our columns are imbalanced.
2. We can see, original length and cleaned length both are imbalanced.
3. By using Wordcloud, we see these words are malignant, highly malignant etc.



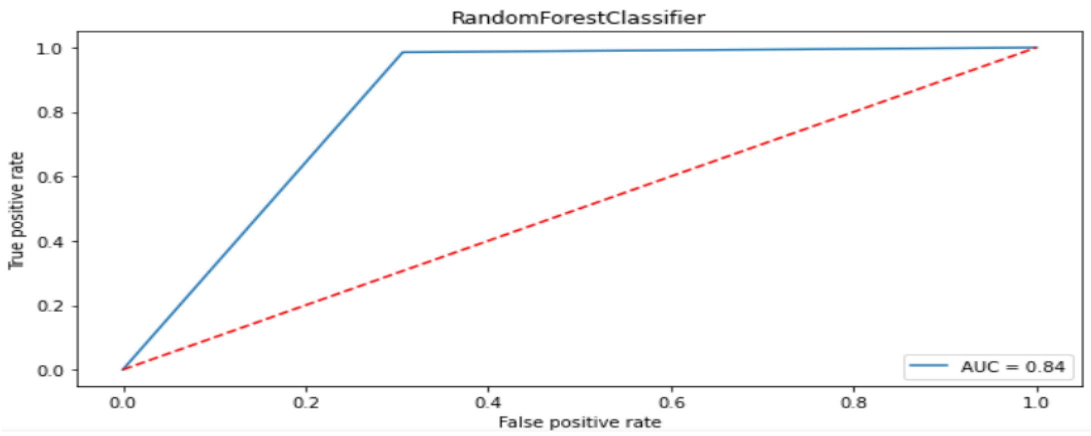
```
log_reg = LogisticRegression() #Mo
dt = DecisionTreeClassifier()
rf= RandomForestClassifier()
knn = KNeighborsClassifier()
svc = SVC()
mb=MultinomialNB()
ada = AdaBoostClassifier()
gbdt= GradientBoostingClassifier()
```

```
RandomForestClassifier()

accuracy_score:  0.956057453688617

cross_val_score:  0.9575048071725663

roc_auc_score:  0.8395390741662518

Hamming_loss: 0.043942546311382946

Log loss : 1.517746454884692
```



```
Classification report:

              precision    recall  f1-score   support

           0       0.84      0.69      0.76      4018
           1       0.97      0.99      0.98     35875

    accuracy                           0.96     39893
   macro avg       0.90      0.84      0.87     39893
weighted avg       0.95      0.96      0.95     39893
```

```
Confusion matrix:

[[ 2787  1231]
 [  522 35353]]
```



1. We check our accuracy by using 6-7 models.
2. Out of which RandomForestClassifier is giving us 96% Accuracy.

CONCLUSION

- Key Findings and Conclusions of the Study:

1. I used various classification methods and out of all machine learning algorithm used, Random forest classifier yields the best results.
2. These malignant comments classification can be used by social media companies to filter and classify some keywords as highly malignant and set their own policy going forward for the customers and other shareholders.
3. For this project I have used wordcloud method that represents the visualization of most frequent words that are highly malignant in nature available in the dataset.

- ## Learning Outcomes of the Study in respect of Data Science:

As per as learning outcomes is concerned, I have learnt the following things in this project:

1. Algorithm need to be used by understanding the dataset for the classification model.
2. From describe method we can get some knowledge related to outliers present in the particular columns (large difference between 75th percentile and maximum percentile)
3. I also understand the visualization of related features and importance related to dataset.
4. I have also used NLTK library to clean the text/comments and find out the actual length of the comments that can be used for further evaluation.

Challenges:

1. It was difficult to load the dataset in notebook as it took some time.
2. Running each line code was a bit slow in notebook, possibly due to low CPU configuration.

· Limitations of this work and Scope for Future Work:

1. Since I have only used a sample dataset, hence sometimes it is difficult to understand the overall impact of this project while filter out the toxic comments in public forum.