PROJECT NAME:-

[MALIGNANT COMMENT PROJECT]

SUBMITTED BY:

SHIVAM SHARMA

INTRODUCTION

Problem Statement:

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.

SO HERE WE HAVE TO-

Build a model which can be used to predict the nature of the comment as there are 6 categories of hated comments which are as follows;-

- **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.

In this case, Label ‘1’ indicates that the comment having hated nature while, Label ‘0’ indicates that the comment is fine …

Review of Literature

THE TOPIC IS ABOUT THE COMMENT WHICH IS SPREAD BY THE USERS THAT ARE ACTIVE ON SOCIAL MEDIA WHERE SOME USERS SENT OFFENSIVE OR HATED COMMENT WHICH CAN DEGRADE ANYBODIES REPOTATION, SO WE ARE BUILDING A MODEL WHICH CAN HELP TO PREDICT THE NATURE OF THE COMMENT MADE BY THE USERS..

BY THE MEANS OF ANALYSING THE COMMENTS IT IS VERY EASY TO PREDICT THE USERS WHICH SPREAD HATE ON SOCIAL MEDIA SO THAT SOME INCIDENT CAN BE PREVENTED.

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….

Data Sources and their format

- **train.csv file downloaded from PMT**

```
data = pd.read_csv('train-malignant.csv')
data.head()
```

	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

- **test.csv file downloaded from PMT**

```
df = pd.read_csv('test-Malignant.csv')
df.head()
```

	id	comment_text
0	00001cee341fdb12	Yo bitch Ja Rule is more succesful then you'll...
1	0000247867823ef7	== From RfC == \n\n The title is fine as it is...
2	00013b17ad220c46	" \n\n == Sources == \n\n * Zawe Ashton on Lap...
3	00017563c3f7919a	:If you have a look back at the source, the in...
4	00017695ad8997eb	I don't anonymously edit articles at all.

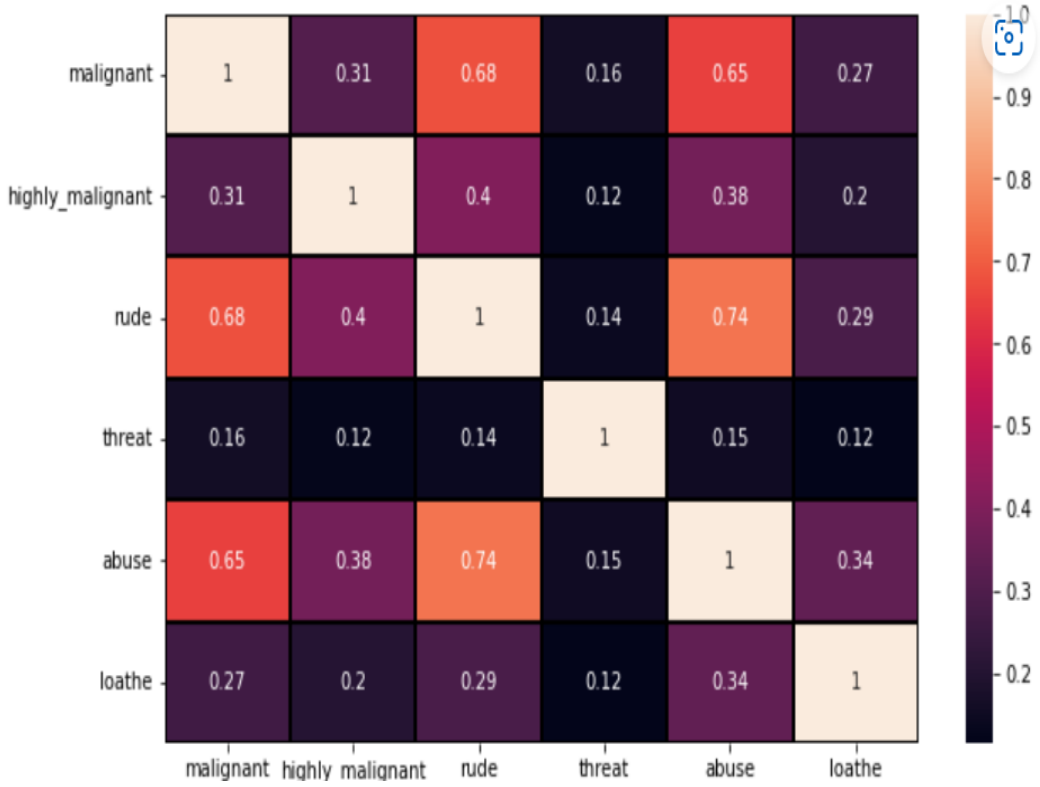
Data Preprocessing and EDA

The steps followed for the cleaning and EDA of the data:-

1) First we analyse the data by simply data.info() and data.describe() methods so that we can check about the datatypes and checked the mean , median and deviation
2) Now we will check the correlation between the independent variables and between the independent variables and dependent variables.
3) Plot a Heatmap to check clearly the correlations between them all.

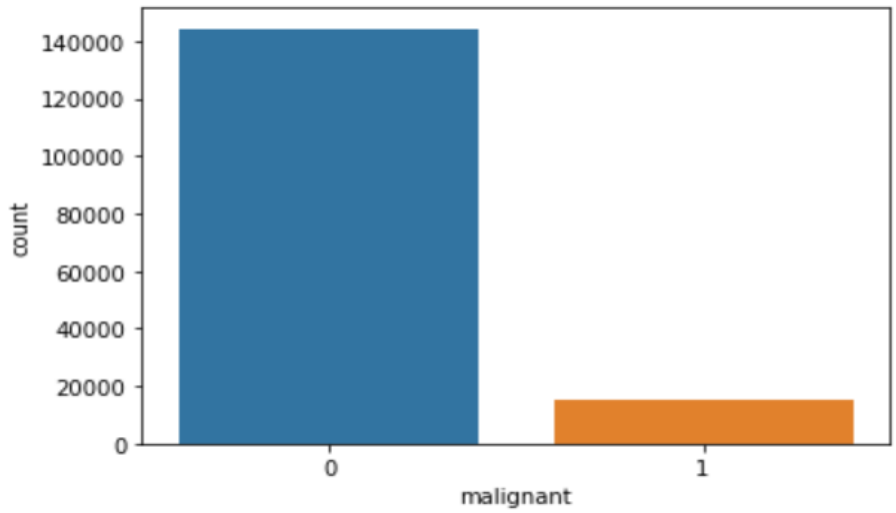
```
data.corr()
plt.figure(figsize=(10,6))
sns.heatmap(data=data.corr(),annot=True,linecolor='black',linewidth=0.2)
```

```
<AxesSubplot:>
```



4) Then after checking the correlation as there is no Multicollinearity present so with the help of Countplot we checked the Count in all target variables I.e Malignant,Highly Malignant,Abuse ,Roathe,Threat and Rude..

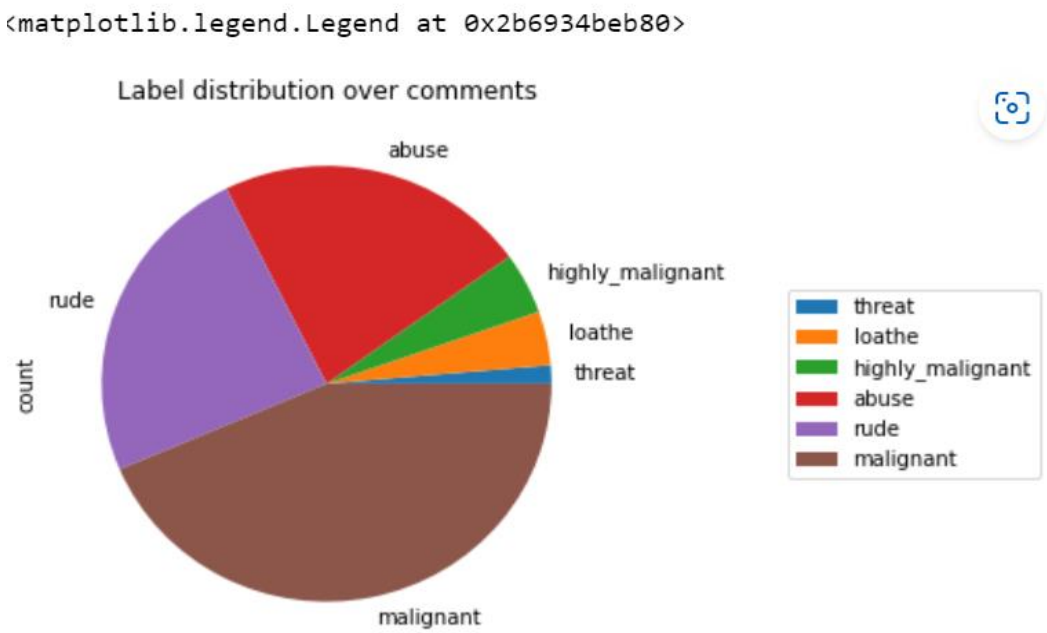
```
malignant


0    144277
1     15294
Name: malignant, dtype: int64
```



- As we can see from the graph the data in the target variable is imbalamced in all targets..

5) Then after we made a new Column of Length which contains the total length of the Comment ..
6) Then after we Convert all comments in lower case, we replaced email addresses with 'email', URLs with 'webaddress', money symbols with 'moneysymb' (£ can by typed with ALT key + 156), 10 digit phone numbers (formats include paranthesis, spaces, no spaces, dashes) with 'phonenumber', and numbers with 'numbr'…
7) Then after we apply lambda function to remove the Stopwords and punctuations from the comment and then we used WordNetLemmatizer so that we can make Lemmas and words can be reduced into their meaningful roots..
8) Then after lemmatizing the comment we made a new column(name as Clear_lenght) so that we can check the original nad clear length separately by doing these cleanings(like stopwords, punctuations and many more as written above)....

9) Then we made a Pie chart for the target varuiables as how they contribute in the comment



- Clearly we can see from the pie chart as the Malignant nature is more in the comments followed by the rude comments..

10) Then after this we made a new column that is “bad” which contains sum of all whole lengths from the all target variables …

11) Then after this we are ready to convert the comments into the vector form as machine can not understand object/strings with the help of TF-IDF Vectorizer…

12) Now we are ready for the model building……

Algorithm used in this project:-

For building machine learning models there are several models present inside the Sklearn module.

Sklearn provides two types of models i.e. regression and classification. Our dataset's target variable is to predict whether fraud is reported or not. So for this kind of problem we use classification models.

But before the model fitting we have to seprate the predictor and target variable, then we pass this variable to the train_test_split method to create the training set and testing set for the model training and prediction.

We can build as many models as we want to compare the accuracy given by these models and to select the best model among them.

I have selected 5 models:

1. Logistic Regression
2. Decision Tree Classifier
3. Random forest Classifier
4. KNN Classifier
5. Gradient boosting Classifier

BEST Algo. From these And why?

Random Forest classifier is the best algo. From all of these algorithms which is used in this data to predict because the difference between the cross val score and the accuracy score is minimum for the Random Forest algo. and it also gives BETTER ACCURACY(approx. 99%)that's why we USED THIS Algo.

- **Lets hypertune this algo with the help of Grid searchCV**
- **Save the model for later predictions**

 # Now my model is ready to predict

CONCLUSIONS

The dataframe contains the data which related to the comment made by the users on social media which contains various kind of nature which can be good or can be offensice/ hatred.
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….

We got our best model i.e RANDOM FOREST CLASSIFIER with the accuracy score of 99%. Here our model predicts 42419 true positive cases out of 42950 positive cases and 3309 true negative cases out of 4922 cases. It predicts 531 false positive cases out of 42950 positive cases and 1613 false negative cases out of 4922 cases. It gives the f1 score of 98%.....

KEYFINDING:-

- As we are not checking the outliers and skewness in this data as there is only 2 column present which contains object datatype and rest are targets..
- As the data is imbalanced we can also used Smote techniques or PSA in this to balance the data..