MALIGNANT COMMENT CLASSIFIER

SUBMITTED BY :
JUHI MISHRA

ACKNOWLEDGEMENT

For completing this project taken help from many of the sources and which are mentioned below :
Kaggle, GeeksforGeeks, , GitHub, Analytics Vidya, Data camp etc. for code or visual presentation help.
My mentor Ms.Khusboo Garg Helped me a lot where I stocked and not able to understand the how to start the project. Problem statement provided by Flip Robo was really helpful in understanding what exactly the problem is. Also other guidance such as two document being provided in which one was how to complete PDF in which format and what are the information to mentioned in detail written in word format, and the other one was what all column means and what are the steps I need to follow to complete this project. Wants to thanks all my mentor and the site from where I got information and able to finish the project on time.

INTRODUCTION

• BUSINESS PROBLEM FRAMING :

Now a days Social media is so popular in every aspect that more than 80% youngsters or even any age people are more aggressively participating in any of the activity either surfing Facebook or Instagram or following any celebrities each and every activity. Positive side if we can see that by connecting more people to these sites business of such companies are in hike even people are coming more closer though they stay far from each other but the darker side is also there which really affects lot many people and leads them to mental trauma

or depression. Likewise we can see that in Instagram there are lot many celebrities posting something related to them on which few really appreciate while few gives such a nonsense and intolerable comments which can't be taken into consideration and sometimes it become so problematic which leads these celebrities or pointed person who being bullied by few of the users to commit sucide or become victim to mental health issues which leads to depression or trauma.With respect to business terms we can say that this type of issues create loss for them, Because lot many avoids being bullied by anyone and skip using these sites who are really active member and very famous. Even few commit sucide or leads to mental trauma because of these comments which is unacceptable , that is also loss for the respective company as there goodwill will be on risk and lot many people will avoid to such sites . So to avoid these issues these particular companies should have eyes on such comments and try to filter those and put them to offensive comments so that others could also understand that these types of comments are not acceptable for anyone . As social media is an open forum and everyone and anyone is allowed to say anything but not in such a way that there one line may hamper the others life.

- **CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM :**

To Understand the problem very first we need to go through the problem statement of the particular problem that we are suppose to work on. Once understood will get an idea what exactly the problem is and how it will help in analysing the same in a proper way . Also what type of problem is this and what approach we need to follow will get an idea once will understand the depth of project. Next using pandas will import the dataset to check how big the dataset is and which supervised machine learning approach do I need to take like regression or classification completely depends upon dataset and target variable. As this dataset was classification based because target column were having binary variables , means whether the comment is offensive or not. Once this is clear we are ready to apply there approaches like data cleaning, preprocessing, feature engineering, and as it is classification based problem so also need to check whether the data is a balanced data or not. Because if data is not a balanced data then our model will perform really bad. As there may be a chances of feature will higher value compare to label or vice versa and model will give more importance to those which will have higher value data. So to balance the same either we oversample or under sample the data depending upon data size. Mainly we focus on oversampling because of no data loss. Data loss may lead our model to perform really bad. Once all these prices done finally we are ready to choose best performing model and based on that will do hyperparamater tuning to improve the score of our model.

- ## REVIEW OF LITERATURE :

Social media is one of the open forum where anyone can give there opinion. However this raised an issue of conflict and hate which in turn making online environments uninviting for users. Though there are lot many platforms where generally people shows there hate in some or other way , but this become so problematic because of lack of model detection online. Such kind of toxic behaviour are much prominent on social media like online hate, abusive language, aggression, cyberbullying, hatefulness and many such other behaviour which express these people hate in different manner. And these type of cyberbullying are increasing day by day such as trolling which many celebrities now a days are facing issues in some or other way . These celebrities have to come across such hateful and offensive comments everyday. Such type of troll and comments may become one of the major reason for mental illness and depression, Which may increase suicidal thoughts. So to diminish such issues we need to build such a model which can predict which comment is offensive and which is hatred comment so that it can be controlled from spreading hatred and cyberbullying . Which means here we need to build such a model which can differentiate between comments and it's categories.

- ## MOTIVATION FOR THE PROBLEM UNDERTAKEN :

Working on such problem statement where I can come across what type of issues going on in society due to cyberbulling and hatred comments, will able to come across mentality of our society and how much is the contribution of new generation who is the future of our country. With this will able to know the mentality of people and can curb such an offensive comments with the help of domain knowledge which can filter such an offensive and hatred comments. None of us wants to be a part of bully or ragging in any of the situation either in school, collage, workplace or within friends . Same thing is for celebrities who are just doing there work but few people are there who are just enjoying by commenting hateful or malignant comments all time. If my domain knowledge can constrain such situation I will really be happy to get them out from such an offensive situation.

ANALYTICAL PROBLEM FRAMING

- ## MATHEMATICAL / ANALYTICAL MODELLING OF THE PROBLEM :

This dataset was not having label column . There were 6 different columns where each comment being marked as 0 and 1 based on review of people on social media.Based on which need to decide if any of the column do have 1 then the target column will have 1 which means that particular comment is a malignant comment and if none of the column were 1 then in target column need to put 0 means comment is not malignant comment. Once analysed and filled the target column accordingly then analysed whether the problem statement is classified or regression based problem. As the target column was having binary variables means 0 and 1 so considering it as classification problem further processed. Regression problem are generally considered when the target column do have continuous variables in it. And here 0 and 1 is not a continuous variables. It's a boolean expression , which means either yes or no. Both train and test data were available so data cleaning and preprocessing required for both of them before model building all the process were same for both train and test data. Total number of rows I training data were 1,59,571 and 8 columns no target column was available here , that we need to create. In test data total number of rows were 1,53,164 and columns were only 2. what type of data are present in problem statement for that used info method and came across that in test data total two columns were having object datatype and in test sheet also two columns were in object datatype as in both ID and comment text were in object datatype. Coming to static analysis none of the columns were in continuous form so the quantile result which was min 0 and max 1 for all the numerical columns even in between which is 1st, 2nd and 3rd quantile 0 was present.There are skewness but no need to treat them as we treat skewness for

continuous variables , because this is one of the mathematical operation which we need to perform for reducing skewness in data. Also there are no null values in both the data sheet. Next created label column by appending 0 for those comment where none of the comments were in 1 and 1 for those where either any of the column were having 1 in it. This dataset is an imbalanced one because 0 is almost having 94% while 1 is only 6% of total data. So to balance this used smote to oversample and avoid loosing data . Mostly all the columns which according to their comment has been categorised 0 was higher in number and with huge margin. Which means there only few people whose comment is unacceptable and offensive. Even by observing the difference between 1 and and 0 in label column we can say that there are 6% people on social media who spread hateful and offensive comments.

- **DATA SOURCES AND THEIR FORMATS :**

This data I received from Flip Robo my intern company both train and test data sheet in csv format. Along with few more required information which were in word and pdf format. Two word sheet been provided in which one was having description of problem statement and the other one was having format detail , means a complete guide how to write detail of project in pdf format. One detail of what are the requirement to complete this project being given. Means from data cleaning to preprocessing , how much model we need to build, detail to be presented in pdf for the project done on malignant comment classifier and at last all to presented beautifully in PPT . Also need to analyse properly using visualisation technique. Loaded both train and test data using pandas in python. So that will able to perform the reared step and can able to finish the project. Snapshot for the same projected below :

```
df_train = pd.read_csv('/Users/juhimishra/Downloads/Malignant Comments Classifier Project/train.csv')
df_train.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

```
df_test = pd.read_csv('/Users/juhimishra/Downloads/Malignant Comments Classifier Project/test.csv')
df_test.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 DONE :

This dataset was on whether the comment of public trolled on anyone is malignant or not where 0 was for not such hatred comment and 1 was for hatred comment. 8 columns were there in this dataset while more than 1.50 lac rows were there. Most of the columns were in numerical form means 0 and 1 as every column were categorised as per the comment done. While there were two column which were in object data type. Those we needed to convert in integer or float form both for train and test dataset as machine algorithm doesn't understand the object or string language. Coming to data cleaning part there were nothing much to do in this as no null values were present in any of the column or rows. Even no columns dropping required as all the columns were equally important in there place. Balancing the data was required as this was classification based problem statement and the data were imbalanced one. Skewness and outliers removal was not required as there were no continuous columns and we generally remove these on continuous column only. Used label encoder to convert object datatype to numerical datatype.

- # DATA INPUTS- LOGIC - OUTUT RELATIONSHIPS :

This dataset was on whether the comment of public trolled on anyone is malignant or not where 0 was for not such hatred comment and 1 was for hatred comment. 8 columns were there in this dataset while more than 1.50 lac

rows were there. There are total 8 input columns which means feature columns on which the target or label column is based. First one was ID column which was used to align the id for each comment. Next is comment column where publicly people commented on social media platform to express there opinion. Next column was malignant with binary values I.e., 0 and 1 depicting which comments are malignant in nature means hostile in nature. After that highly malignant option was there which was one step ahead of malignant. Then next category was rude which means which column was considered to be rude . Then coming to next category which was threat means which column was considered as threat. Next was abuse where people used really very bad language or we can slang for any particular person. Loathe was the last feature which means hateful comment done on anyone. These were the few features based on which we prepared a target column considering any of the columns if having 1 that will be considered as 1 which means a particular comment is hateful and offensive in nature and if none of the column is 1 then in target column putted 0 which means comment is not offensive in nature. With this come across that our target was completely dependent on other columns and was having clear relationship among them. Always all the feature column is considered as independent one while target column is a dependent column. If any of the independent column is not proper of target will column will be affected directly.

- **HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED :**

To build a good model we should have a cleaned dataset and if not then need to work on same so that will get best possible outcome. Data Processing and data cleaning are one of the vital job role of data scientist. In this problem . Hardware tools which being used while coding
were :

1. CPU - This is one of the important part while performing machine learning task because most of the computation will be done in learning

environment which is most likely done on CPU.

2. Power Supply - This is also one of the important hardware while performing any machine learning task as if power supply will not be

there , will not able to perform the task.

3. RAM - While performing coding on machine learning minimum 8 GB ram is required to perform the task.

Coming to software part the important libraries and packages required was for machine learning algorithm :

1. Pandas - One of the important library which helps in importing the data to perform data manipulation and analysis.

2. Numpy - This library is used to perform all the mathematical operations required during coding

3. Seaborn and Matplotlib - Used for visualisation analysis. As we know that to understand any data in a more proper way and for better presentation, visualisation is one of the important technique used.

MODEL/S DEVELOPMENT AND EVALUATION

- **IDENTIFICATION OF POSSIBLE PROBLEM - SOLVING APPROACHES :**

For any supervised machine learning algorithm first we need to decide whether the provided problem statement is a regression based or classification based problem statement and this we decide by target column where if data present inside is continuous in nature then it's a regression based problem statement and if the data inside is in binary or discrete form then it's a classification problem. Here were having binary based target column where 0 and 1 was present as a result based on all feature where 0 means that a particular comment is not malignant in nature while 1 means that a particular comment is malignant in nature. So as a data scientist we need build such a model which can differentiate between comments and their categories. Based on what type of dataset do I have start analysing what are the preprocessing steps do I need. To follow. Because preprocessing is one of the important steps before model building. First identified what are the datatypes do I have , and here except two rest all 6 columns were in integer form. So we need to use any encoding technique as per the requirement to convert these string datatype to integer or float datatype. Because machine algorithm doesn't understand the object data type and to make them understand we need to convert object column to integer or float datatype. Then checked missing values of any present in dataset . There were no null values so no need to use mean, mode or replace method here. Coming to static part here most of the columns were in binary form means either the comment is hatred or not. Even skewness and outliers can't be used here because of no continuous columns present in this dataset. As this was classification based problem statement so needed to check whether the particular dataset do have imbalanced data or not. Checked the same using count method and found that almost 0 was 94% and 1 6% only which means the dataset was clearly an imbalanced one and to balance them used smote method for oversampling so that will not loose any of the data. Once all the steps of preprocessing done scaled all the features bring them to single unit and start building model for prediction purpose and to know which model performs best for this problem statement.

- **TESTING OF IDENTIFIED APPROACHES (ALGORITHMS) :**

Before choosing any algorithm first we need to identify which type of problem we are working on, means regression or classification. As based on that we select our model for prediction purpose. And here I choose total 5 algorithm to check which performs best based on which will do hyperparamater tuning to improve score if possible. Below listed the name of all algorithm selected for this problem statement :

1. Decision Tree Classifier
2. Random Forest Classifier

3. Support Vector Classifier
4. Logistic regression (because this was binary based problem statement)
5. XG Boost

And for Hyperparameter tuning used GridSearchCV

- **RUN AND EVALUATE SELECTED MODELS :**

Before choosing any algorithm first we need to identify which type of problem we are working on, means regression or classification. As based on that we select our model for prediction purpose. And here I choose total 5 algorithm to check . First I found the best random state based on which will train my model. Using decision tree classifier found best random state which was 421 and the accuracy score on this random state was 96.56% . Below is the snapshot for the same :

```
maxaccu = 0
maxrs = 0

for i in range(1,500):
    x1_train,x1_test,y1_train,y1_test = train_test_split(x_scaler,y1,test_size = 0.30,random_state = i)
    dt = DecisionTreeClassifier()
    dt.fit(x1_train,y1_train)
    pred = dt.predict(x1_test)
    acc = accuracy_score(y1_test,pred)
    if acc>maxaccu:
        maxaccu=acc
        maxrs=i
print("Best Accuracy score is:",maxaccu,"On Random state: ",maxrs)
```

```
Best Accuracy score is: 0.9656311040833411 On Random state:  421
```

After finding best random state trained the model on the same by distributing the ration 70:30 which 70% of data assigned for training the model based on which model will able to understand the previous data and will predict accordingly.

```
x1_train,x1_test,y1_train,y1_test=train_test_split(x_scaler,y1,random_state=i,test_size=0.30)
```

Very first algorithm which I build for this problem statement was Decision Tree Classifier model . Snapshot for the same given below :

```
DTC = DecisionTreeClassifier()
DTC.fit(x1_train,y1_train)
pred = DTC.predict(x1_test)
acc = classification_report(y1_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           0       0.96      0.97      0.96     42954
           1       0.97      0.96      0.96     43054

    accuracy                           0.96     86008
   macro avg       0.96      0.96      0.96     86008
weighted avg       0.96      0.96      0.96     86008
```

```
print(cross_val_score(DTC,x_scaler,y1,cv=5).mean())
```

```
0.9637973638950765
```

Above we can see that accuracy score based on this model understanding was 96% while precision score was 97% and recall and f1 score is 96%. There is not much difference of score in any of the report. Also checked cross validation score based on 5 folding and it was 96.37% which means not a huge difference between two score which is accuracy and CV score only 0.37% was the difference.

Then used Random Forest algorithm to build a model. Screenshot for the same mentioned below:

```
RFC = RandomForestClassifier()
RFC.fit(x1_train,y1_train)
pred =RFC.predict(x1_test)
acc = classification_report(y1_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           0       0.97      0.96      0.96     42954
           1       0.96      0.97      0.96     43054

    accuracy                           0.96     86008
   macro avg       0.96      0.96      0.96     86008
weighted avg       0.96      0.96      0.96     86008
```

```
print(cross_val_score(RFC,x_scaler,y1,cv=6).mean())
```

```
0.963560197005846
```

Random Forest score is similar to Decision Tree Classifier score . In both the accuracy score is 96% and CV score only 0.2% difference is there which is very minor.

Then used SVC algorithm . Snapshot for the same given below :

```
SV = SVC()
SV.fit(x1_train,y1_train)
pred = SV.predict(x1_test)
acc = classification_report(y1_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97     42954
           1       1.00      0.95      0.97     43054

    accuracy                           0.97     86008
   macro avg       0.98      0.97      0.97     86008
weighted avg       0.98      0.97      0.97     86008
```

```
print(cross_val_score(SV,x_scaler,y1,cv=6).mean())
```

```
0.974198791734684
```

Compare to both above model SVC performance is better with 97% accuracy score and precision was almost 100% while recall is 95% and with that we can see the affect on f1 score which is based on both precision and recall is 97%. Though the difference is not much but model performance is 1% better than DTC or RFC score.

Then used Logistic regression. This algorithm generally we use when do have binary based target column and a classification problem statement. Below is the snapshot for the same:

```
LR = LogisticRegression()
LR.fit(x1_train,y1_train)
pred = LR.predict(x1_test)
acc = classification_report(y1_test,pred)
print(acc)
```

```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97     42954
           1       1.00      0.95      0.97     43054

    accuracy                           0.97     86008
   macro avg       0.98      0.97      0.97     86008
weighted avg       0.98      0.97      0.97     86008
```

```
print(cross_val_score(LR,x_scaler,y1,cv=6).mean())
```

```
0.974198791734684
```

Score for logistic regression and SVC are almost same which is 97% of accuracy score and Precision is 100% while recall is 95% and f1 score is 97%. CV score is also same as I did on 6 folding for both SVC and logistic regression. Will decide later on which I need to do hyperparameter tuning.

Next used XgBoost for model building. Snapshot for the same mentioned below :

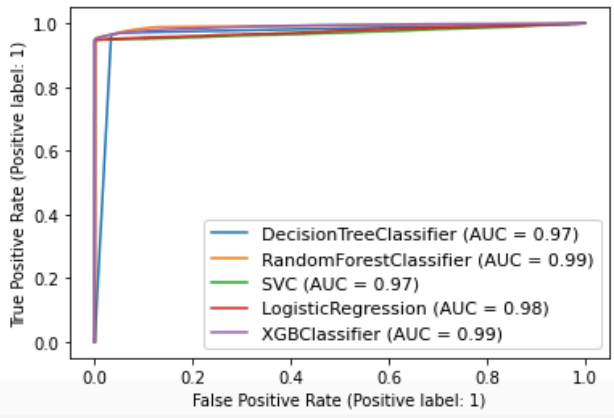
```
# xgboost
import xgboost
xgb = xgboost.XGBClassifier()
xgb.fit(x1_train, y1_train)
pred = xgb.predict(x1_test)
acc = classification_report(y1_test,pred)
print(acc)
```

```
[09:20:33] WARNING: /opt/concourse/worker/volumes/live/7a2b9f41-3287-451b-6691-43e9a6c0910f/volume/xgboost-split_1619
728204606/work/src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective
'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
behavior.
              precision    recall  f1-score   support

           0       0.95      1.00      0.97     42954
           1       1.00      0.95      0.97     43054

    accuracy                           0.97     86008
   macro avg       0.98      0.97      0.97     86008
weighted avg       0.98      0.97      0.97     86008
```

Cv score is 97.42%. Which means all the three algorithm that is Logistic regression, SVC and XgBoost all are performing almost same with same accuracy score which is 97% and Cv score which is 97.42%. Here all the model performance are almost same with not much difference between there score. So I decided to run hyperparameter tuning on SVC and check how much improvement do I get on score after tuning the model. Using GridSearchCv tuned the model.

Once selected and runner all the algorithm plotted the same on AOC RUC Curve :

```
disp = plot_roc_curve(DTC,x1_test,y1_test)

plot_roc_curve(RFC,x1_test,y1_test,ax = disp.ax_) # ax = axes with confusion mtrix

plot_roc_curve(SV,x1_test,y1_test,ax = disp.ax_)

plot_roc_curve(LR,x1_test,y1_test,ax = disp.ax_)

plot_roc_curve(xgb,x1_test,y1_test,ax = disp.ax_)

plt.legend(prop={'size':11}, loc = 'lower right')

plt.show()
```



As from above graph we can see that almost all the model are on same level both on false positive rate and true positive rate. There is a minor difference Random Forest and XGBoost are having 99% AUC score. Here True positive means which we considered as true and in reality that particular comment is considered as true only while false positive means which we considered to be true but in reality it was false.

- **KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION :**

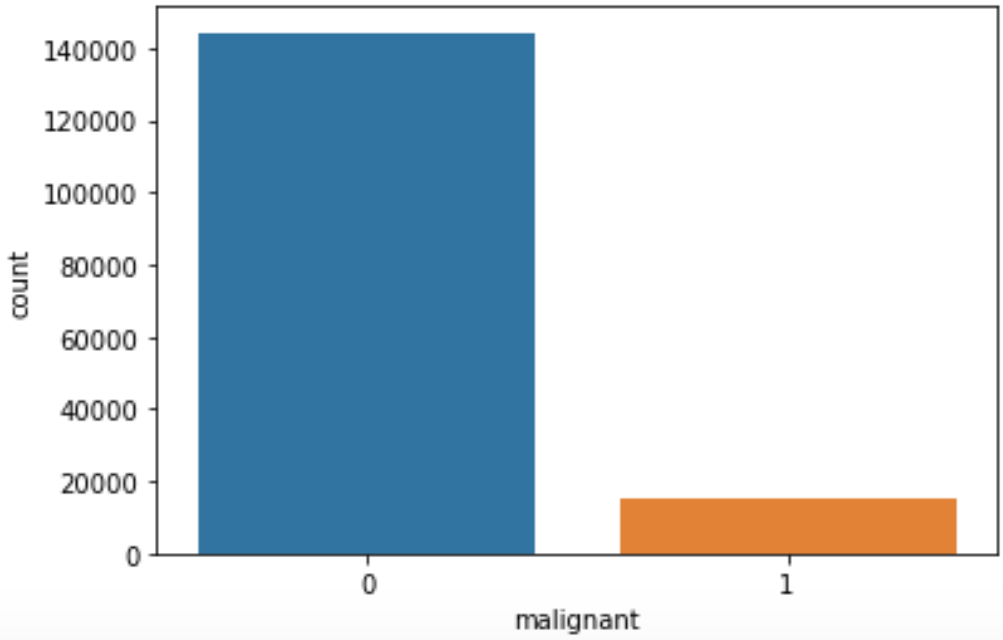
Before choosing any algorithm first we need to identify which type of problem we are working on, means regression or classification. As based on that we select our model for prediction purpose. And here I choose total 5 algorithm to check . To solve any problem statement first we need to know what type of problem is? Means as we know that we working on supervised machine learning but in that too there are two different types which are : Regression Problem and classification problem. Regression problem generally we use when our target column is in continuous nature while classification problem we use when the target column do have either binary form data in it or discrete at a which means classifying something. This problem statement where I need to predict whether the comment is malignant as per it's category or not is a classification problem statement as target column do have binary data in it which is 1 and 0. After identifying the problem statement analysed the data once loaded to python using pandas . Then checked how big the data is which means shape of row and column in which rows were more than 1.50 lac both in train and test dataset while columns were only 8 before target column. Once target column created than in train dataset total number of columns were 9 while in test dataset only 2 column which is id and comment column. Now we need to build a good model based on which will predict the required result.And for having good model we need to do proper data cleaning and preprocessing. These are the few factors which are important to get success in best model performance.

- **VISUALISATION :**

Visualisation is one of the important aspect to analyse any data. Visualisation helps in understanding the data more deeper and clearer, If anyone who doesn't able to understand by text they can even understand by going through graph and plot. Here for getting visualisation analysis import seabron and matplotlib library. As there not much column and mostly we need to analyse here comment and it's categories . Here we were having comment column based on which categories been given like malignant, High malignant , rude, threat, abuse, loathe. According to comment category been decided and created label based on all other category. Below visualised few of the columns using count method.

```
sns.countplot('malignant',data=df_train)
```

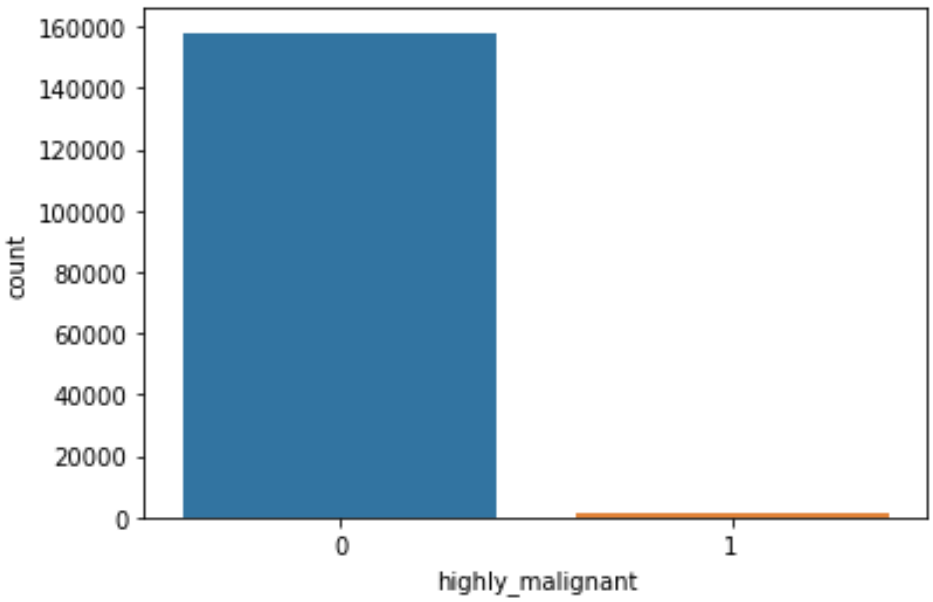
```
<AxesSubplot:xlabel='malignant', ylabel='count'>
```



Above using count method we can see that 0 which means non malignant comment is much higher than malignant comment which means there are only few people who do these type of comments which being considered as malignant comment.

```
sns.countplot('highly_malignant',data=df_train)
```

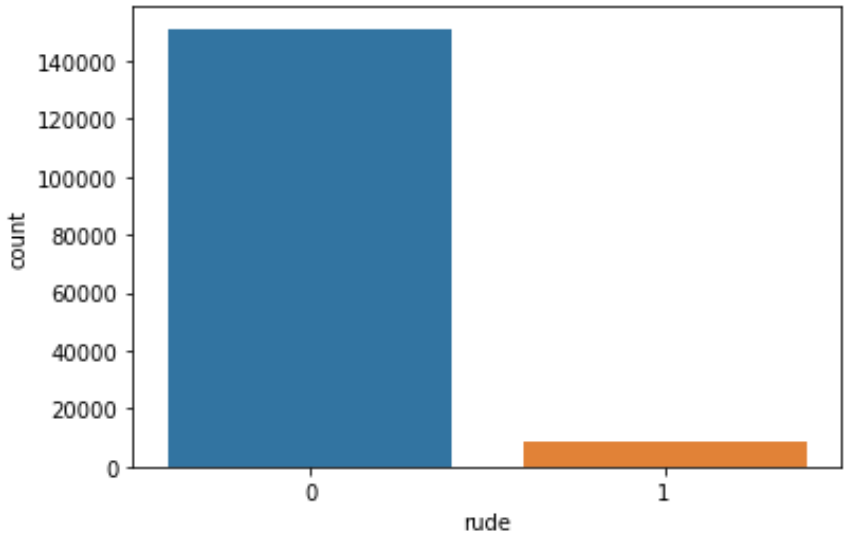
```
<AxesSubplot:xlabel='highly_malignant', ylabel='count'>
```



There are only 1% people whose comments are considered as highly malignant comment. Highly malignant means hateful or offensive comments.

```
sns.countplot('rude',data=df_train)
```

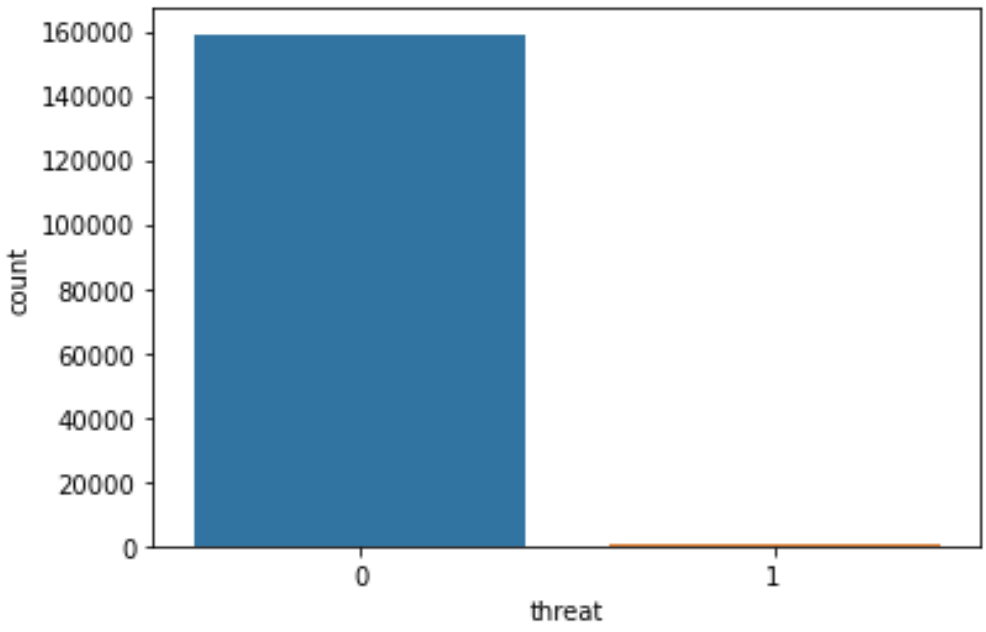
```
<AxesSubplot:xlabel='rude', ylabel='count'>
```



Here also in the above plot we can see that rude comments are almost below 20K which means there are few people who generally uses such a comment against anyone which is non acceptable.

```
sns.countplot('threat',data=df_train)
```

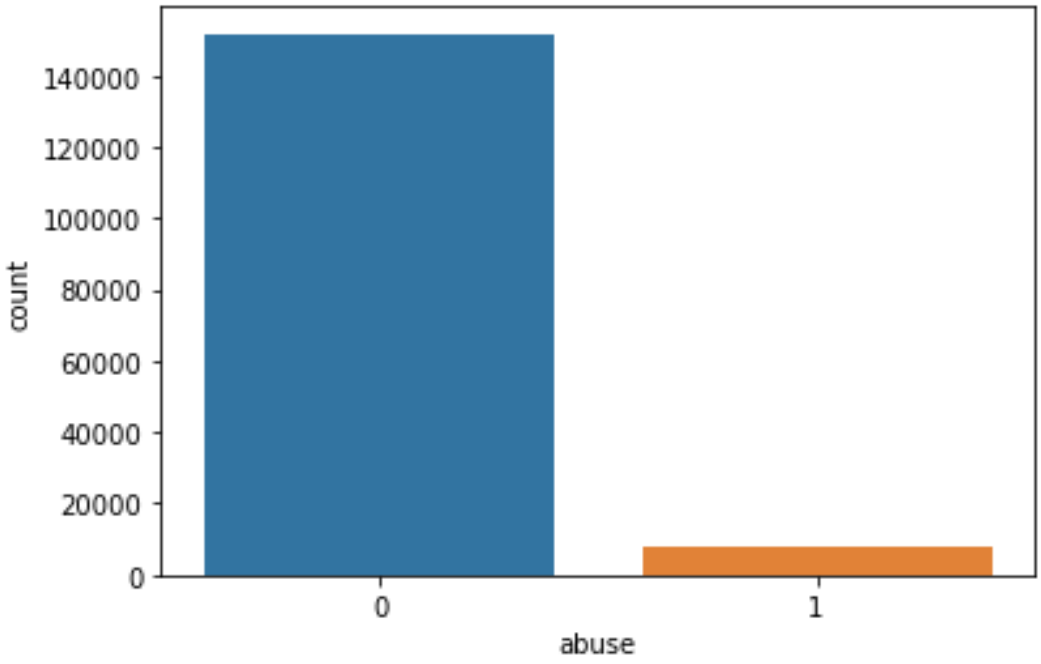
```
<AxesSubplot:xlabel='threat', ylabel='count'>
```



Threat comments are almost negligible which even we can say in percentage compare to non threat comments are in 0.10% . That means we can assume generally none of the people uses such a threaten comments for others on social media.

```
sns.countplot('abuse',data=df_train)
```

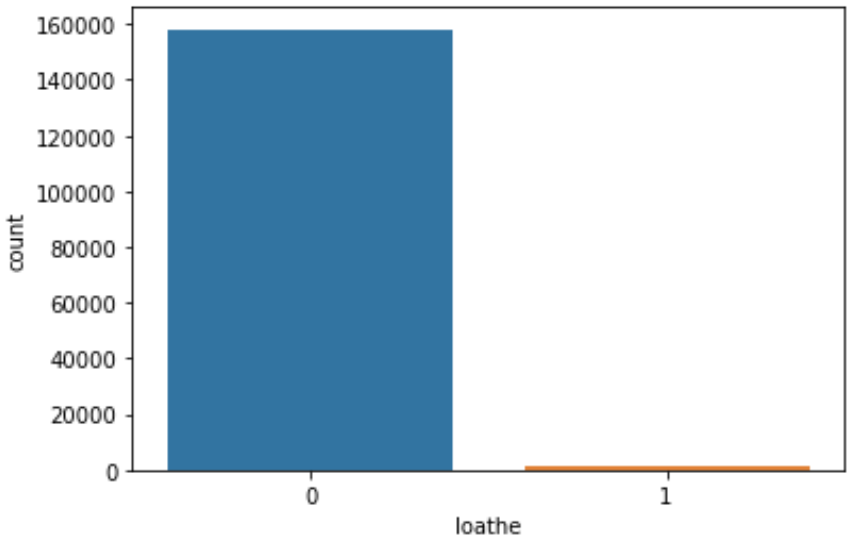
```
<AxesSubplot:xlabel='abuse', ylabel='count'>
```



Abusive comments are more compare to threat comments. Here we can say that due to hateful nature or to hurt someone most of the people uses such comments. To curb such users we need to detect them and try to mute them so that t may not affect the other person on whom the comment being done.

```
sns.countplot('loathe',data=df_train)
```

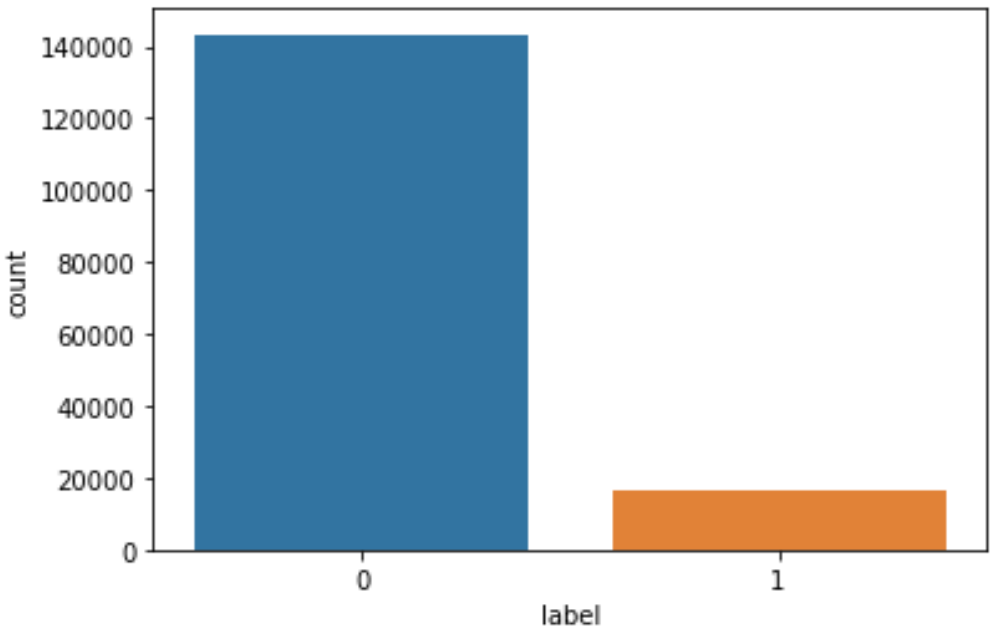
```
<AxesSubplot:xlabel='loathe', ylabel='count'>
```



Loathe which also means hateful or more dislike comments done on someone. These types of comments generally done on social media in troll mode . But these type of comments sometimes affects someone so much that it may bring social thought too within them. Here loathe comments are too less compare to abusive or malignant comments.

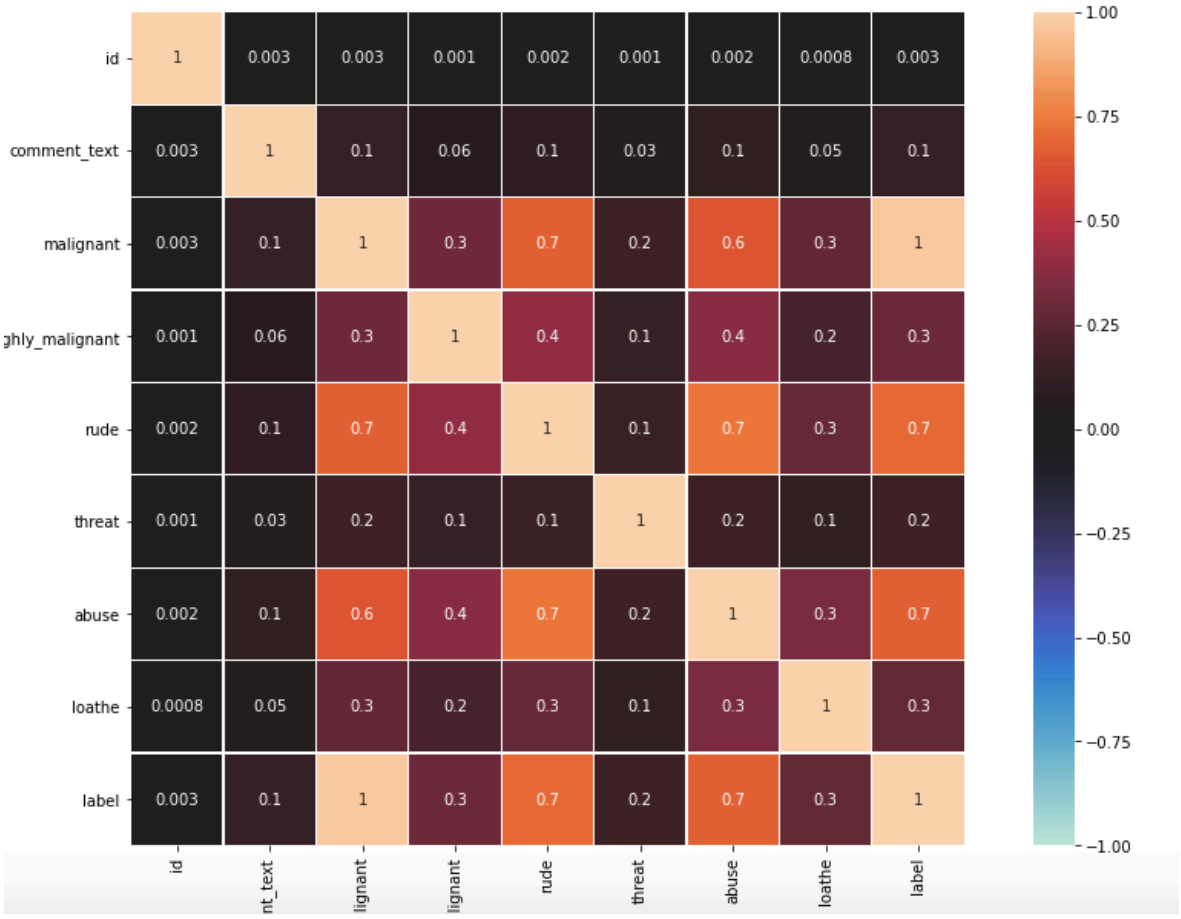
```
sns.countplot('label',data=df_train)
```

```
<AxesSubplot:xlabel='label', ylabel='count'>
```



Label is my target column where I appended from all other categories 0 and 1 where 0 is for non hateful comment and 1 for hateful comment. After appending all got the result as overall there 6% population who mainly enjoys trolling others. That we need to curb using our domain knowledge.

Using heat map tried to analyse the relation between features and also features to label. If any of the feature is highly correlated than there will be a chance of overfitting. So to avoid that we need to delete those features which are highly correlated to each other



None of the features are highly correlated to each other as we can see using heat map above. Threat is only the column which shows relation from features and even with label. But not that much that we need to delete the column.

- # INTERPRETATION OF THE RESULT :

With the help of visualisation come across the insight of data to have clear view of our problem statement. Once if we understand the problem statement then it become easy to decide further what steps we can take to build a good model. There were few columns which were categorised based on comments. And did

analysis on those category columns so that will able to understand which type of abusiveness being used more by people in public forum. As a data scientist my job role is to understand the data properly and build such a model which can differentiate between comments and it's categories. Which means need to predict which type of comments are more hateful. Above in visualisation we observed that threat and loathe comments being used very less compare to malignant, rude, abusive. But overall there are 6% people who do such type of trolling which may affect others life on whom the comment being done. Also by using heat map observed the correlation between features . There are relation but not that much that we need to delete that particular column. Highly correlated features means they both have same function and we need to select one which is having higher value. Next coming to preprocessing and data cleaning here no null values were present in this dataset so no need to treat for the same. Next there were two columns which were ID and comment column which we need to convert to numerical form using encoding method as these column were in object datatype and our machine algorithm doesn't understand the string datatype. Here I used label encoder to convert string datatype to integer form. Next checked whether the data is an imbalanced data or not as this is a classification problem. And this dataset was an imbalanced dataset. Using smote did oversampling of data so that will not loose data . After doing all preprocessing, feature engineering and data cleaning scales all the features after splitting them to respective variables x and y . Using standard scaler did scaling so that all the features comes to same scale and model performance will be much better as it can able to understand the data in a proper way. Once scaling done find best random state and trained the model by using train test split. 70% data assigned for training purpose so that my model will be trained on 70% data and for test given 30% on the basis of which my model will predict.
Then used 5 classification algorithms to get the accuracy and CV score. Based on best performing model did hyperparameter tuning using GridSearchCV and at last did prediction based on test dataset.

CONCLUSION

- ## KEY FINDINGS AND CONCLUSION OF THE STUDY :

Internet now a days become part of every individual ans specially social media. Social media is such a platform where most of the population spends there time and try to know the life of others, either it's a friend , family member, or any celebrities. Here every one free to share there opinion or can put their point of view what they think for others. There is another part which really affects others specially celebs who every day have to be trolled by public. Sometimes few comments are tolerable while some comments are such which really affects their life. These comments affects so much that few of them face the issue of depression, mental illness and even social thought comes to few minds. These hateful, abusive comments should be filtered or to be put into trash so that others life not get affected. Here as a data scientist need to build such a model which can differentiate between comments and it's categories. To build a model need to collect data related to these comments and need to analyse the data properly so that we can train the model and can predict what result being expected. Once both train and test data collected in the form of CSV format loaded the same to python using pandas in read format. Checked the shape of dataset where in train data total columns were 1,59,571 and columns without label were 8 and after adding label which is our target column were 9, while in test data total rows were 1,53,164 and columns were 2 in which only id for comments and comments column were there and both were in object datatype. In train dataset also comments and id column were in object datatype while all the there category column were in numerical form. Converted object datatype to numerical form using label encoder technique. Machine algorithm doesn't understand the text to string datatype , to make them understand we need to convert string to integer or float datatype. No null values were present in this dataset. Coming to skewness and outliers no need to remove though skewness were there in numerical columns but as we know that skewness generally we perform for mathematical operation and these operation we do on continuous columns and none of the columns were in continuous form here. Columns which were in numerical form were discrete in nature. Comments being segregated into 6 categories which is malignant, highly malignant, rude, abusive, threat, and loathe. According to the comments done 1 and 0 value being given respectively i.e., 0 for non hateful comments and 1 for hateful comments. Where using visualisation technique we can observe that hateful comments ratio are too

less compare to good comments where threat category were almost nill. All together there were 6% population who do such kind of comments. And we need to filter these only. As we can see in our target 0 and 1 was given means either yes or no which means it's in binary form. So this is classification dataset and in classification problem we need to check whether the dataset is balanced or not. Checked using count method and found that there is a huge difference between the value of 0 and 1. So to balance the same used smote method to oversample the data and to avoid data loss. As data loss lead our model not to perform well.Once all the preprocessing done scaled all the feature using standard scaler to bring all the unit to same label. Next found best random state so that will train our model on the same. Segregated 70% and 30% respectively for training and testing purpose . Which means my model will be trained on 70% data and will predict on the basis of 30% data. Then choose 5 models which were Decision Tree Classifier, Random Forest Classifier, SVC, Logistic Regression and XGBoost. Among all almost three model were having same performance and given highest accuracy and CV score which was 97% . Based on score did hyperparameter tuning for SVC algorithm using GridSearchCV. Then saved the model using job lib. And predicted on the basis of test dataset.

- **LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE :**

Internet now a days become part of every individual ans specially social media. Social media is such a platform where most of the population spends there time and try to know the life of others, either it's a friend , family member, or any . To understand any of the data first we need to go through the whole dataset thoroughly , which sometimes become boring if we need to go theoretically . Visualisation is one of the technique using which we make any dataset interesting and more easy to understand. Here also using seaborn and matplotlib two of the libraries which is used to represent visualisation technique tried to get insight of data. There were category given based on comments done. And as we observed that there very few people who given threat comments. Mostly people used rude, malignant and abusive comments. All total there were 6% population who used such type of comments. Coming to data cleaning part , here I need to convert few of the columns to integer or float form as they were in object form and as there were no null values so no need to do anything for that. This dataset is a classification based problem statement so need to check whether it's a balanced one or not as if the data is not a balanced data model will not perform well and prediction will be affected. Once preprocessing done scaled all the feature to bring them to same unit so that our

model will able to understand the dataset properly. Then found best random state and selected 5 algorithms which were Decision Tree Classifier, Random Forest Classifier, SVC, Logistic regression and XGBoost. As the dataset was big so running each model was taking too much time and doing hyperparameter tuning are way longer. That was the main issue what I faced mostly. Restarted kernel and run the output for 2 - 3 times so that the page will be refreshed and I can get the result asap.

- ## LIMITATION OF THIS WORK AND SCOPE FOR THE FUTURE WORK :

There were not much to do in this dataset , means data cleaning part was too less , no null values, no skewness removal or outliers as there were no continuous columns. Less to explore in this dataset even for visualisation part too. So coming to learning part there were very few steps to do. Future score I can say is this is one type of dataset on which if we are comfortable will help in understanding lot many such social media platform which are now a days are very popular. Lot many job scopes are there as there are lot many companies who do have a eyes on what comments they receive on there site and in what term. They can decide better why how to improve there work on the basis of same and can filter those which is affecting others life.