FLIP ROBO

MALIGNANT

COMMENT CLASSIFIER

PROJECT

Submitted by:

Akash chaudhary

Intenship 17

ACKNOWLEDGMENT

The project malignant comment prediction is a problem with celebrities pages comments on the behalf of the different kind of people who commented on the celebrities pictures and their videos. The data frame was given by My mentor Mr. keshav bansal and help me a lot to address the problems and solutions.

INTRODUCTION

Business Problem Framing

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

With the covid 19 impact in the market, we have seen lot of changes in the online markets. After the covid 19 we have seen drastically change in the numbers of the internet users and social media users hence the viewers and followers of the celebrities also increased but the problem is everyone is not their fan. Some users give malignant comment on their pages. 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 Collection Phase

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.

• Review of Literature

Every celebrity have their own social media site or page or channel people are free to like share or comment on their content so this comments are either good or some are obviously bad so we need to make a project to find out the bad comments so we have make the project to found the malignant or bad comments using machine learning algorithms. 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.

- Motivation for the Problem Undertaken

As of time is changing digitalisation is also increasing everyone now uses social platform and social media every most of the people have their own social media handle and today the number of social media users are increasing day by day. Everyone use to surfing the social sites celebrities pages they are also free to comment on their pages or their content so these comment are either good or bad so we have make the model to find out the malignant or rude comment by using machine learning algorithms this make my analytical skills strong and the market can demand this type of analysis. It also enhance my personal skills. To make this type of project and being the perfectionist in such type of work could help the companies and celebrities agents and other people to getting the idea of the malignant or bad comment comment.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

The mathematical problems are included are this was the huge data so getting null values is not a big deal. So there is some null values we need to tackle with them because of huge data there are many rows i.e. 159000 rows first of all we have understand the data than we see the data types and what we need to do. Checking for outliers adjust them and some other works are accured at the time of making the model and collecting the data through web using web driver. We used random forest regressor to predict the bad comment and putting them into the test data to finalize the result.

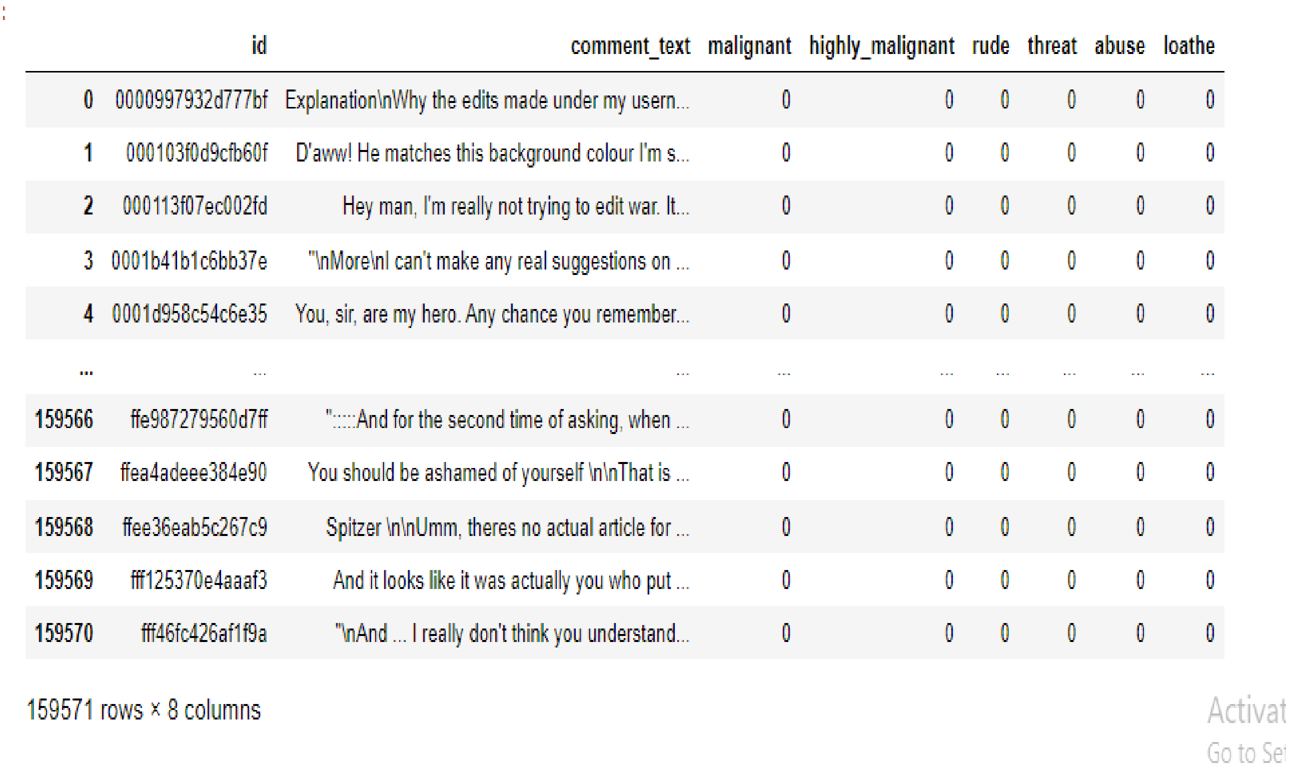
Data Sources and their formats

The data is gathered from the celebrities social media pages project was assigned to me by mr.keshav bansal.

The data was gathered from celebrities social media channel for making the machine learning model to classify the bad comment . The data contains 159000 rows and 8 columns. The data includes both in integer and string forms.

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.



	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
...	...	...	...	...	...	...	...	...
159566	ffe987279560d7ff	":::::And for the second time of asking, when ...	0	0	0	0	0	0
159567	ffea4adeee384e90	You should be ashamed of yourself \n\nThat is ...	0	0	0	0	0	0
159568	ffee36eab5c267c9	Spitzer \n\nUmm, theres no actual article for ...	0	0	0	0	0	0
159569	fff125370e4aaaf3	And it looks like it was actually you who put ...	0	0	0	0	0	0
159570	fff46fc426af1f9a	"\nAnd ... I really don't think you understand...	0	0	0	0	0	0

159571 rows × 8 columns

• Data Preprocessing Done

First of all we got data the data in csv file form so we import this into our workbook. we need to clean up the data although its almost cleaned because we get the sorted data so we just added only those columns which we need the most so first we need to clean the data in the data there are 159000 rows and 8 columns. So first we check the null values in the columns so we have distribute the bad and normal comments in different parts and we got 89.83% normal comments and 10.17% bad comments. So we have check the data distribution of the bad comments and their categories Than we correlate the targeted variable with some important independent variables. For data cleaning we lower all the comments and remove some artificial words or numbers like nan, /n, umm etc. by using nltk than we make the word cloud to find the bad words and classify accourdingly.

• Data Inputs- Logic- Output Relationships

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.

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

Presumptions are includes the data was only for celebrities youtube video comments because there was not any like or anything which are used for other social media channels like instagram or facebook.

- **Hardware and Software Requirements and Tools Used**

We have used jupyter notebook to make the model and imported some python libraries these are includes numpy, pandas, maths, stats, seaborn, matplotlib, sklearn etc.

Model/s Development and Evaluation

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

It was a regression problem that's why we have used logistic regression. We have used TfidfVectorizer to scale the data and also balancing of the data. We have used decision tree classifier because it gives as the best accuracy score and its cross validation score was also good fit to the test set. And the difference between accuracy and cross validation score is very less. So we took this and also done the hyper parameter tunning to tune the data. Also we have find out the AUC ROC score and curve. Which represents the accuracy.

- ## Testing of Identified Approaches (Algorithms)

Algorithms which are used for training and testing are:

Logistic regression

Ada boost regressor

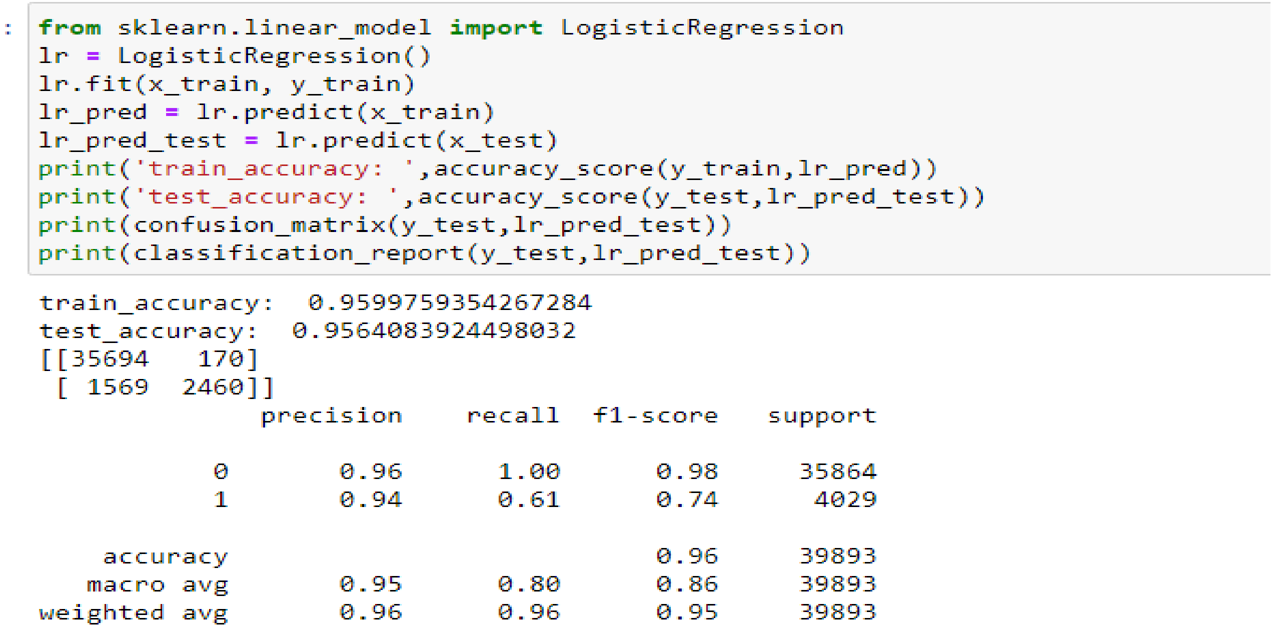
Decision tree regressor

Random forest regressor.

- ## Run and Evaluate selected models

There are the snapshots of the models that we used:

1. Logistic regressor



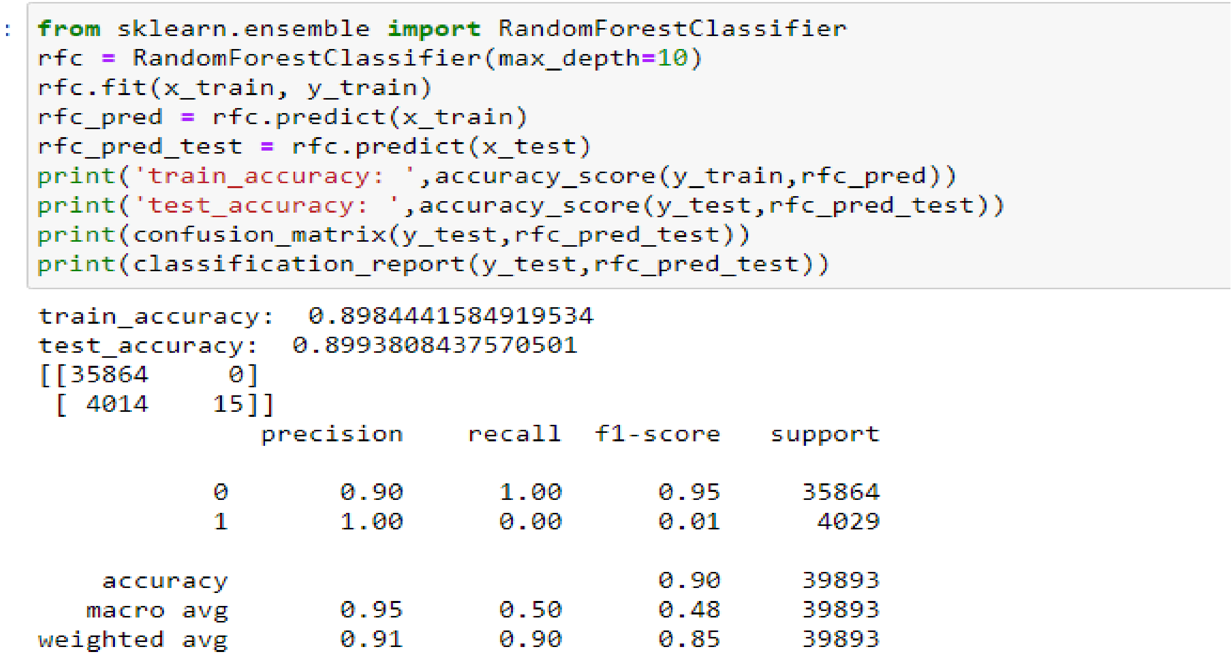
```
: from sklearn.linear_model import LogisticRegression
  lr = LogisticRegression()
  lr.fit(x_train, y_train)
  lr_pred = lr.predict(x_train)
  lr_pred_test = lr.predict(x_test)
  print('train_accuracy: ',accuracy_score(y_train,lr_pred))
  print('test_accuracy: ',accuracy_score(y_test,lr_pred_test))
  print(confusion_matrix(y_test,lr_pred_test))
  print(classification_report(y_test,lr_pred_test))
```

```
train_accuracy:  0.9599759354267284
test_accuracy:  0.9564083924498032
[[35694   170]
 [ 1569  2460]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     35864
           1       0.94      0.61      0.74      4029

    accuracy                           0.96     39893
   macro avg       0.95      0.80      0.86     39893
weighted avg       0.96      0.96      0.95     39893
```

2. Random forest regressor.



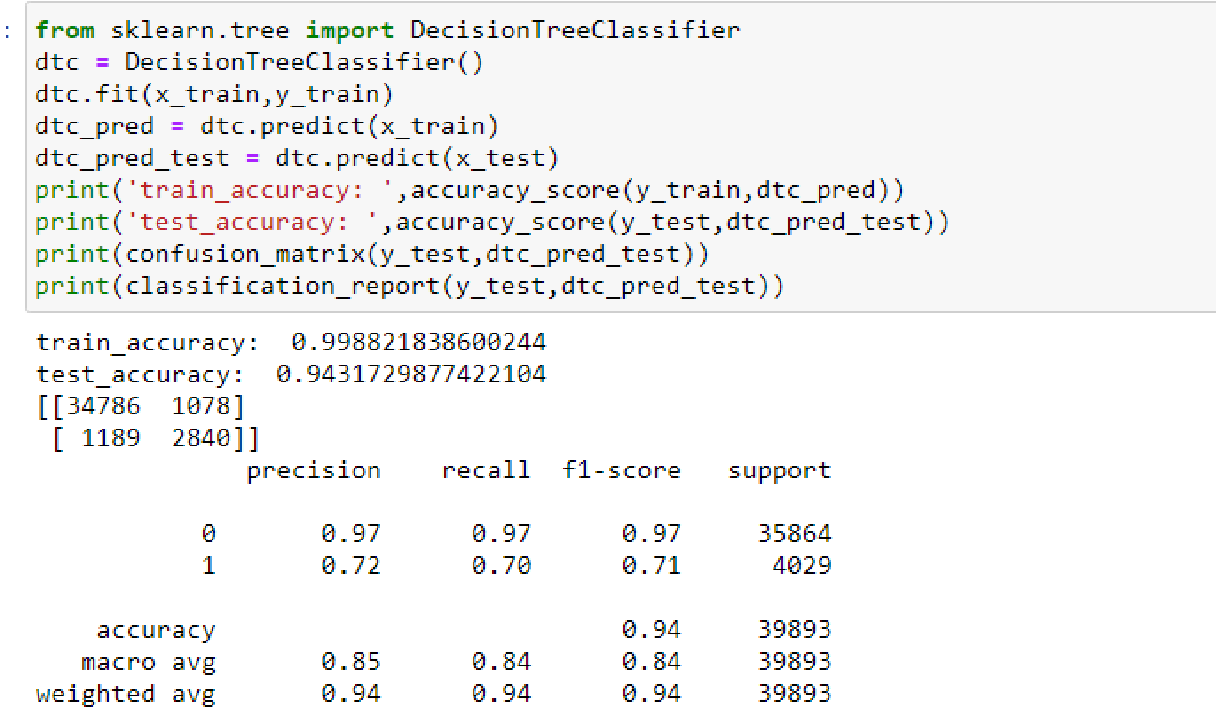
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(max_depth=10)
rfc.fit(x_train, y_train)
rfc_pred = rfc.predict(x_train)
rfc_pred_test = rfc.predict(x_test)
print('train_accuracy: ',accuracy_score(y_train,rfc_pred))
print('test_accuracy: ',accuracy_score(y_test,rfc_pred_test))
print(confusion_matrix(y_test,rfc_pred_test))
print(classification_report(y_test,rfc_pred_test))
```

```
train_accuracy:  0.8984441584919534
test_accuracy:  0.8993808437570501
[[35864     0]
 [ 4014    15]]
              precision    recall  f1-score   support

           0       0.90      1.00      0.95     35864
           1       1.00      0.00      0.01      4029

    accuracy                           0.90     39893
   macro avg       0.95      0.50      0.48     39893
weighted avg       0.91      0.90      0.85     39893
```

3. Decision tree regressor



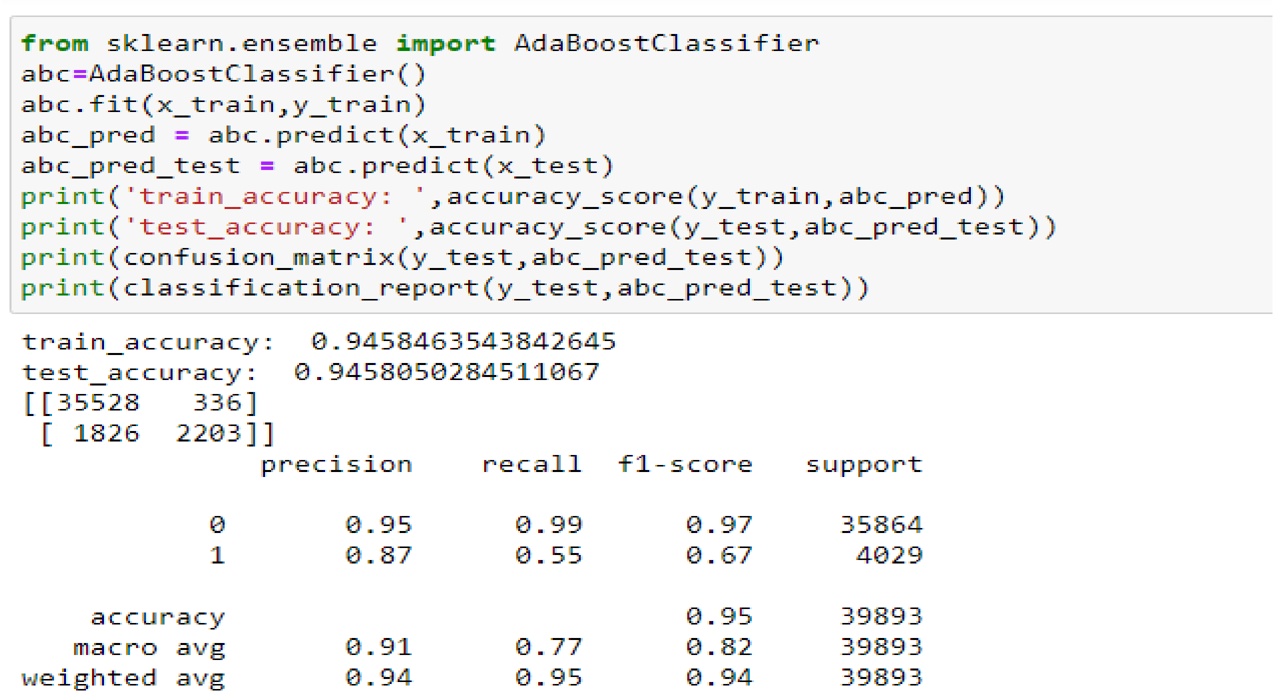
```
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train,y_train)
dtc_pred = dtc.predict(x_train)
dtc_pred_test = dtc.predict(x_test)
print('train_accuracy: ',accuracy_score(y_train,dtc_pred))
print('test_accuracy: ',accuracy_score(y_test,dtc_pred_test))
print(confusion_matrix(y_test,dtc_pred_test))
print(classification_report(y_test,dtc_pred_test))
```

```
train_accuracy:  0.998821838600244
test_accuracy:  0.9431729877422104
[[34786  1078]
 [ 1189  2840]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     35864
           1       0.72      0.70      0.71      4029

    accuracy                           0.94     39893
   macro avg       0.85      0.84      0.84     39893
weighted avg       0.94      0.94      0.94     39893
```

4. Ada boost regressor



```
from sklearn.ensemble import AdaBoostClassifier
abc=AdaBoostClassifier()
abc.fit(x_train,y_train)
abc_pred = abc.predict(x_train)
abc_pred_test = abc.predict(x_test)
print('train_accuracy: ',accuracy_score(y_train,abc_pred))
print('test_accuracy: ',accuracy_score(y_test,abc_pred_test))
print(confusion_matrix(y_test,abc_pred_test))
print(classification_report(y_test,abc_pred_test))
```

```
train_accuracy:  0.9458463543842645
test_accuracy:  0.9458050284511067
[[35528   336]
 [ 1826  2203]]
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     35864
           1       0.87      0.55      0.67      4029

    accuracy                           0.95     39893
   macro avg       0.91      0.77      0.82     39893
weighted avg       0.94      0.95      0.94     39893
```

So we have chose the decision tree regressor because of its best fit score, accuracy score, best test accuracy and cross validation score.

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

The key metrics which are used are:

Train and Test accuracy score

Confusion matrix

Classification report

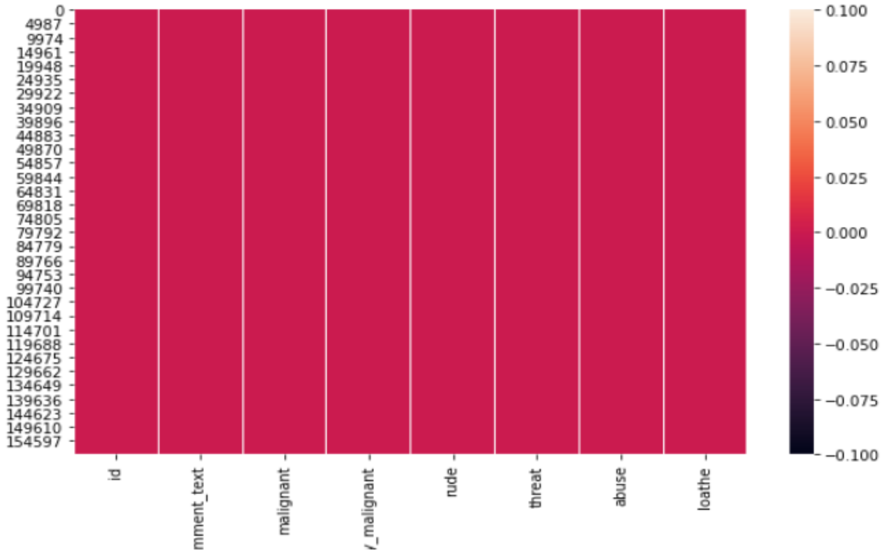
Cross validation score

We check the confusion matrix which includes precision, recall, f1 score and support. And also checked the cross validation score because the difference between the fit score and cross validation score indicates models accuracy.

- **Visualizations and Interpretation of the Results**

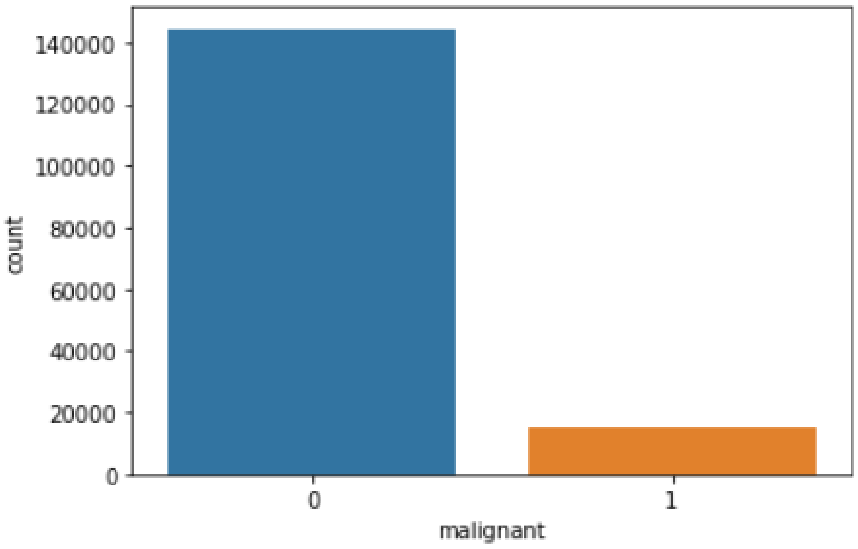
The visualizations are includes:

1. Heat map for null values



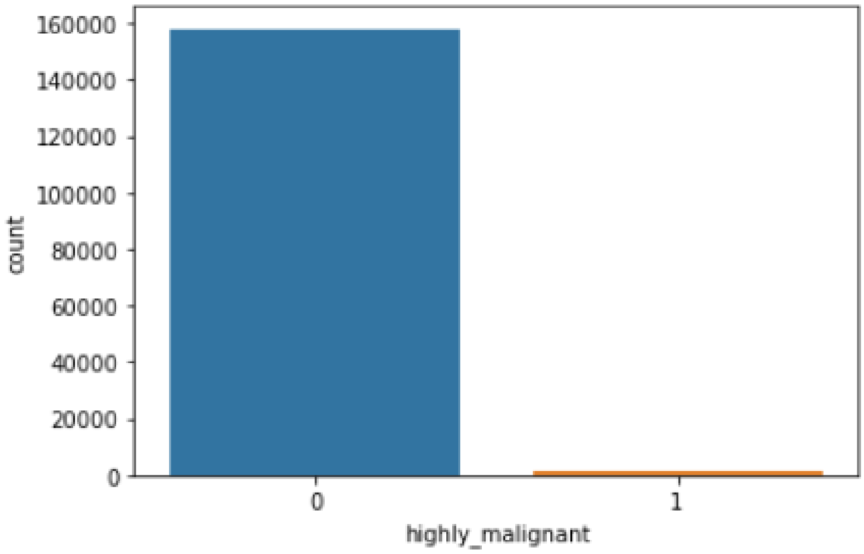
the dataset is free of null values.

2. Countplot for malignant comments.



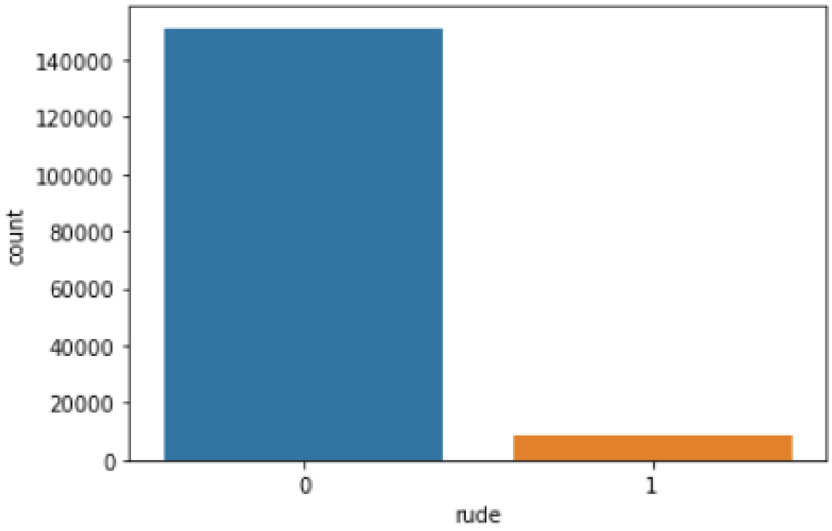
The number of malignant comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that tha comment is malignant or not.

3. Count plot for highly malignant comments.



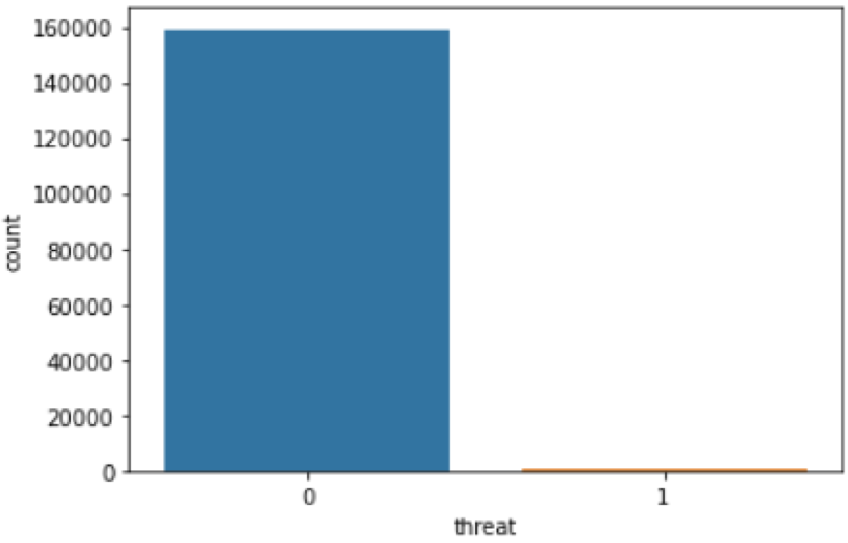
The number of highly malignant comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that tha comment is malignant or not.

4. Countplot for rude comments.



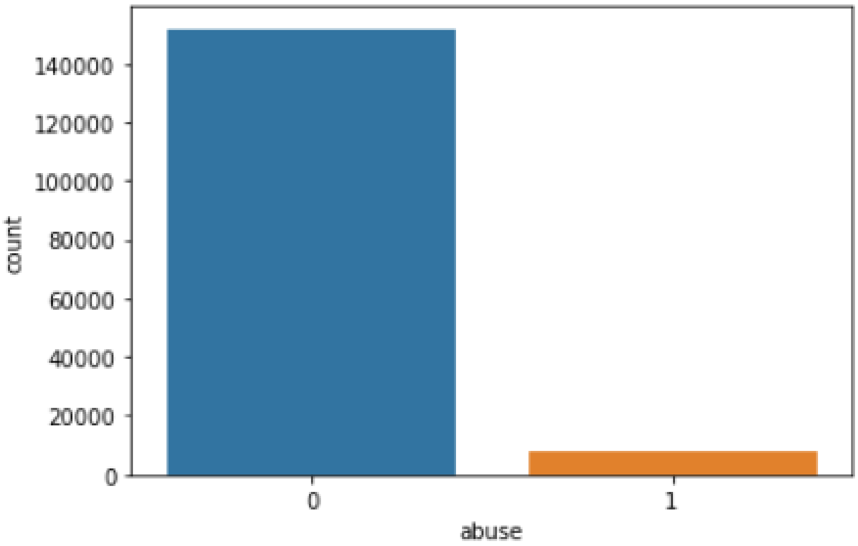
The number of rude comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that tha comment is rude or not.

5. Countplot for threat full comments.



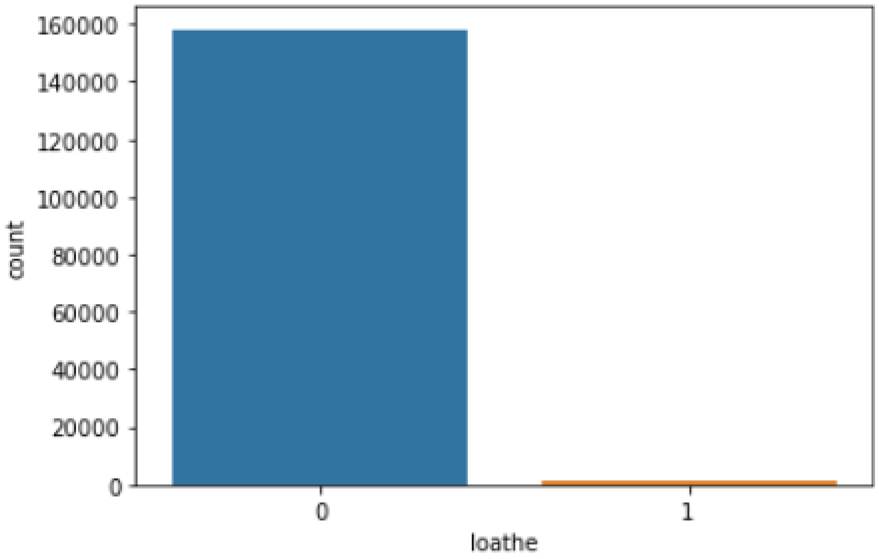
The number of threat comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that the comment is threat or not.

6. Countplot for abuse comments.



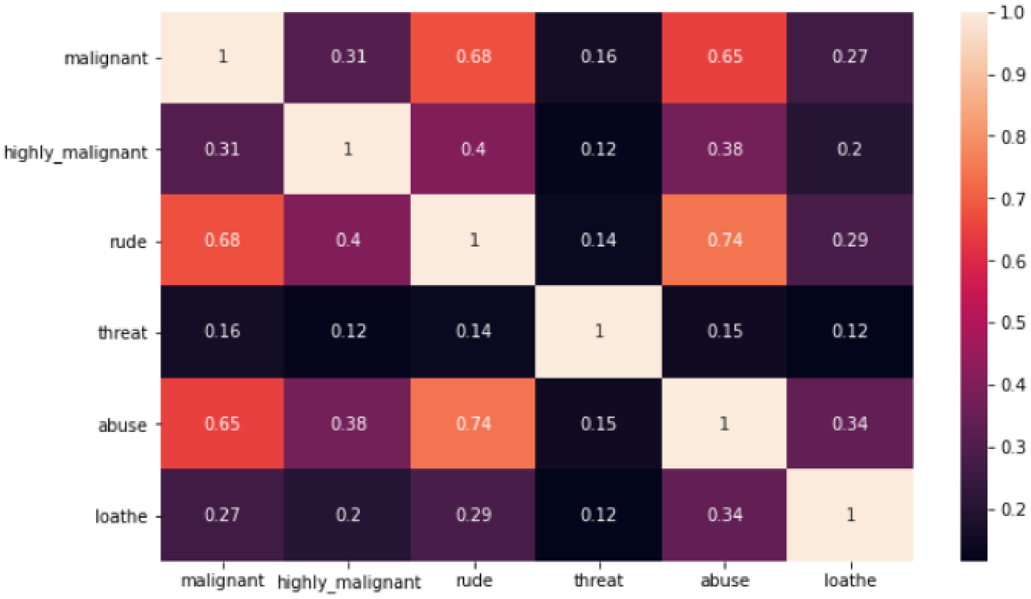
The number of abusive comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that the comment is abusive or not.

7. Countplot for loathe comments.



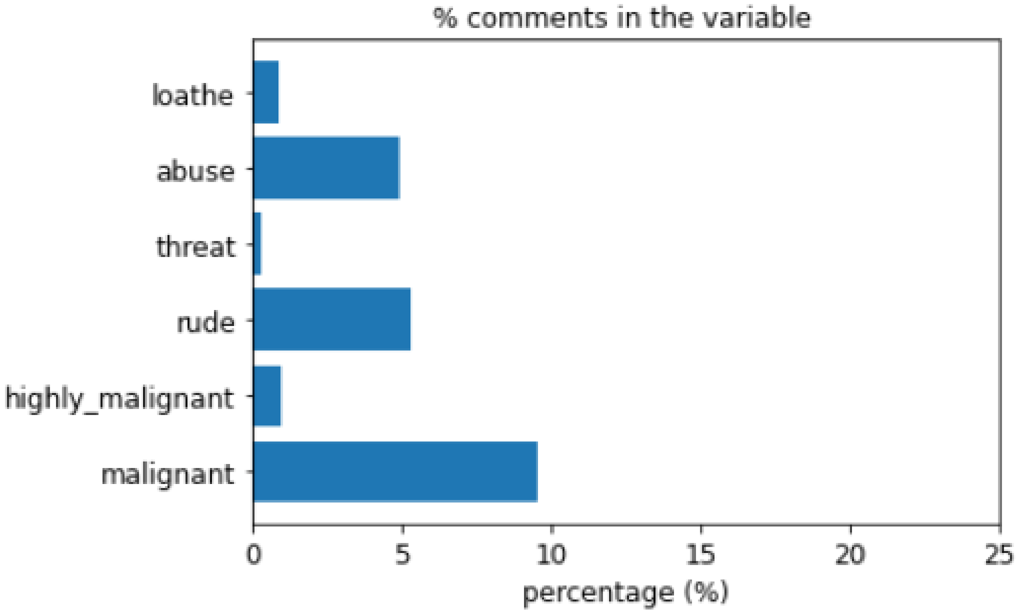
The number of loathe comments are very low as compare to normal comments. Here the blue bar represents no and 1 represents yes that the comment is loathe or not.

8. Correlation matrix for vairables



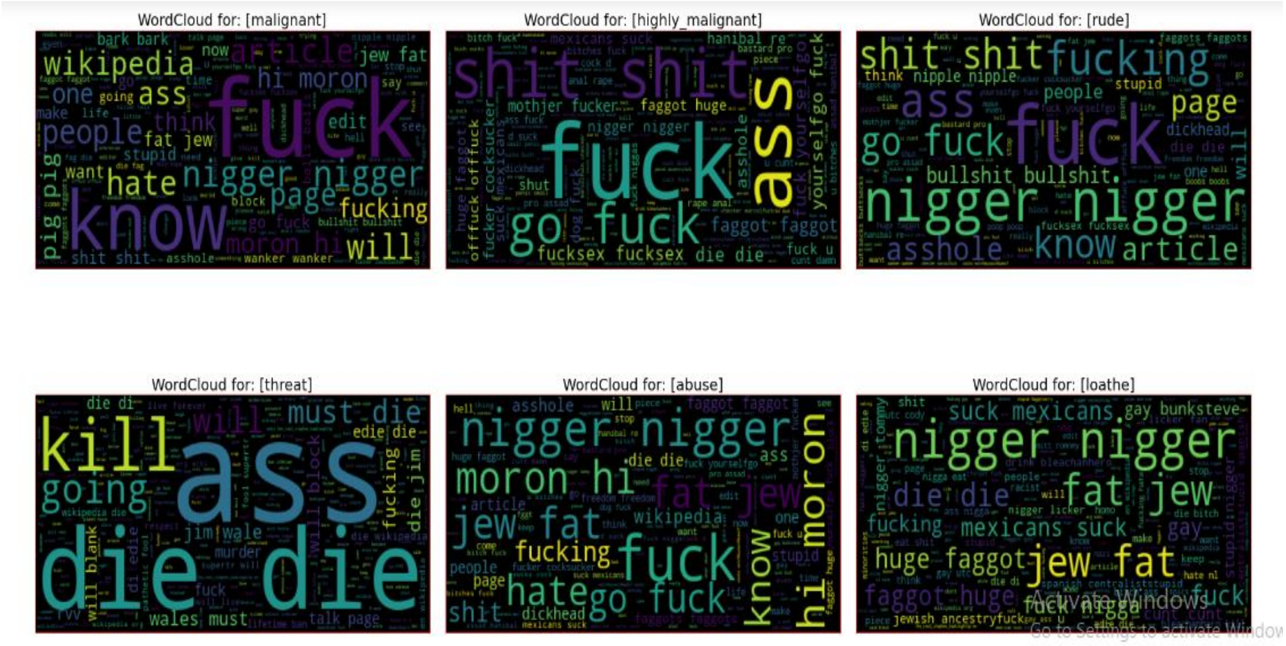
variables are not much correlated with each other

9. % distribution of all the comments categories.



We can see we have maximum numbers of malignant comments than rude than abusive very fer percentage comments of highly malignant, loathe and threat.

10. Display the bad comments in all the categories.



now we can see the mostly used words in the malignant categories are includes fuck, fucking, nigger, stupid etc.

now we can see the mostly used words in the highly malignant categories are includes fuck, ass, shit, fucksex etc.

now we can see the mostly used words in the rude categories are includes nigger, bullshit, dickhead etc.

now we can see the mostly used words in the threat categories are includes ass, die, kill etc.

now we can see the mostly used words in the abuse categories are includes nigger, moron, go fuck, jew etc.

now we can see the mostly used words in the loathe categories are includes jew, fat, nigger, nigga etc.

CONCLUSION

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

As we already know the time is changing digitalisation is also increasing everyone now uses social platform and social media every most of the people have their own social media handle and today the number of social media users are increasing day by day. Everyone use to surfing the social sites celebrities pages they are also free to comment on their pages or their content so these comment are either good or bad so we have make the model to find out the malignant or rude comment by using machine learning algorithms this make my analytical skills strong and the market can demand this type of analysis. It also enhance my personal skills. To make this type of project and being the perfectionist in such type of work could help the companies and celebrities agents and other people to getting the idea of the malignant or bad comment comment.

The conclusion is includes many factor as I mentioned earlier could effect the mental health of the celebrities or any individual. Malignant comments or threatfull comments and abusive comments on someone page is not right thing. And because of these the user didn’t focus on useful comments so we used the machine learning algorithm to make a project which classify the bad comments with the 98% accuracy in our model now if user wants to remove the bad comments he/she can filter it out and read only the good once.

• Learning Outcomes of the Study in respect of Data Science

Because the data is huge the number of rows and columns are also high so these make my understanding of data increased. I have tried many things to clean up the data than find which variables are used for prediction and make the model powerful I have correlate many useful independent variables with the targeted variable so that I can understand the data more. After visualisation I have correlate the data with target in numbers to understand which independent variables gives how much impact on the targeted variable. So after all of these I used TfidfVectorizer to scale the data after scalling the data train the model than used the different machine learning models. And get the best score in Decision tree classifier than doing hyper parameter tuning etc. these steps improve my machine learning and model building skill

• Limitations of this work and Scope for Future Work

Because we get the good accuracy score and best fit score we have choose the best model. We can assume the estimated classifications are almost correct. But we our model does not gives us 100% accuracy and many other factors like real time commenting some words which we did not get by using the wordcloud are also includes. So we can took the estimation by using these model but also need to see the market trend.