***Malignant-Comments-Classifier Project***



***Submitted By:***

***Harish Dhami***

***(Internship:Flip Robo Techknology@Bangalore)***

***Acknowledgements***

The completion of this thesis study would not have been possible without the support and extensive knowledge of several website.

***Introduction***

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.

***Analytical***

***Problem Framing***

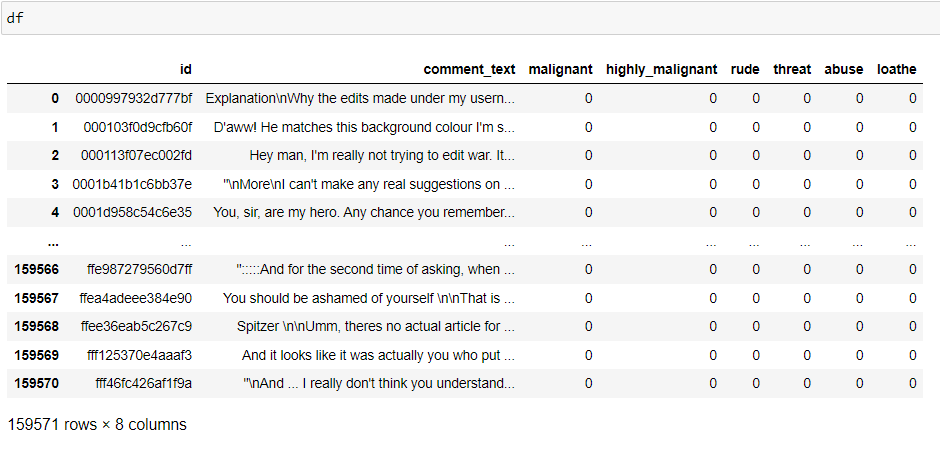
***Mathematical/ Analytical Modeling of the Problem:-*** In this project, we will develop and evaluate the performance and predictability of trained and tested models based on comments which is provide by flip robo techknology. Once we get a good fit, we will apply on our test data.

In here we will use various classification algorithm to predict our target. Let's have an overview of the algorithms we will use for our predictions. To read more about these algorithms , just click on the algorithms name.

* [LogisticRegression](https://www.google.com/search?q=linear+regression&rlz=1C1CHBF_enIN997IN998&oq=&aqs=chrome.1.69i59i450l8.734952339j1j15&sourceid=chrome&ie=UTF-8):- Logistic regression analysis is valuable for predicting the likelihood of an event. It helps determine the probabilities between any two classes. In a nutshell, by looking at historical data, logistic regression can predict whether: An email is a spam.
* [DecisionTreeClassifier](https://www.google.com/search?q=about+DecisionTreeRegressor&rlz=1C1CHBF_enIN997IN998&ei=7kG5YoWNM6fA3LUPqcGy8AQ&ved=0ahUKEwiFvLry9Mz4AhUnILcAHamgDE4Q4dUDCA4&uact=5&oq=about+DecisionTreeRegressor&gs_lcp=Cgdnd3Mtd2l6EAM6BAgAEA1KBAhBGABKBAhGGABQAFjfEWDdFWgAcAF4AIABqQKIAZYLkgEDMi02mAEAoAEBwAEB&sclient=gws-wiz):- Decision trees **help you to evaluate your options**. Decision Trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options.
* [SVR](https://www.google.com/search?q=about+SVR&rlz=1C1CHBF_enIN997IN998&oq=about+SVR&aqs=chrome..69i57j0i10i22i30l6j0i390l3.4767j1j15&sourceid=chrome&ie=UTF-8):- The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value.
* [KNeighborsClassifier](https://www.google.com/search?q=about+KNeighborsRegressor&rlz=1C1CHBF_enIN997IN998&oq=about+KNeighborsRegressor&aqs=chrome..69i57j33i160.3952j1j15&sourceid=chrome&ie=UTF-8):- By default, the KNeighborsClassifier looks for the 5 nearest neighbors. We must explicitly tell the classifier to use Euclidean distance for determining the proximity between neighboring points. Using our newly trained model, we predict whether a tumor is benign or not given its mean compactness and area..
* [RandomForestClassifier](https://www.google.com/search?q=about+RandomForestRegressor&rlz=1C1CHBF_enIN997IN998&ei=n0a5Yq5xxJWx4w_O07lA&ved=0ahUKEwjuvN-u-cz4AhXESmwGHc5pDggQ4dUDCA4&uact=5&oq=about+RandomForestRegressor&gs_lcp=Cgdnd3Mtd2l6EAMyBwghEAoQoAEyBwghEAoQoAE6BwgAEEcQsAM6CggAEOQCELADGAE6BQgAEIAEOggIABCxAxCDAToICAAQgAQQsQM6CwgAEIAEELEDEIMBOgUIABCRAjoLCC4QgAQQsQMQgwFKBAhBGABKBAhGGAFQywhYjiBgrzJoAXABeAKAAbQDiAH5EJIBCTAuMi4yLjAuM5gBAKABAaABAsgBDcABAdoBBggBEAEYCQ&sclient=gws-wiz):- What is Randomforestclassifier in Python?A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting..

***Data Sources and their formats:-*** The dataset which I use for model making is provide by FlipRobo Techknology, The data set has 159571 rows and 8 columns

Dataset looks as follows:-



***About The DataSet:-***

|  |  |
| --- | --- |
| **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. |

***Data Preprocessing Done:-***

For the purpose of the project the dataset has been preprocessed as follows:

* Checking shape of the dataframe
* Checking Missing Value
* Checking which type of data stored in each columns
* Text processing
* Plot Word cloud
* Visualization
* Describing the dataset
* Checking correlation and using heatmap for better understanding

We’ll now open a python 3 Jupyter Notebook and execute the following code snippet to load the dataset and remove the non-essential features. Recieving a success message if the actions were correclty performed.

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

* ***Hardware:- Desktop/Laptop***
* ***Software:- Anaconda***
* ***Libraries:-*** ***Numpy,Pandas,Matplot,Seaborn,nltk,etc***

***Model/s Development and Evaluation***

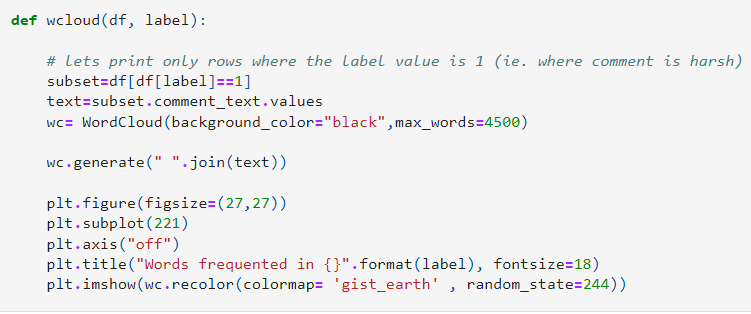
* ***Identification of possible problem-solving approaches (methods):***
* ***Missing Value Handling:***

**There are 2 primary ways of handling missing values:**

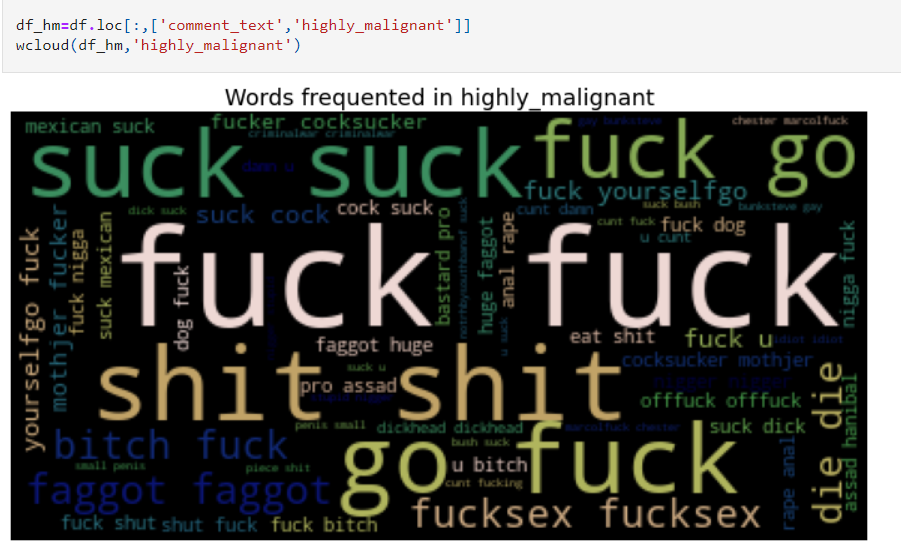
* Deleting the Missing values:-Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.
* Imputing the Missing Values:- There are different ways of replacing the missing values
* Replacing With Mean
* Replacing With Mode
* Replacing With Median,etc.
* We are free from missing value otherwise it is very important step for model building
* ***Data Input Output Logic with WordCloud:-***

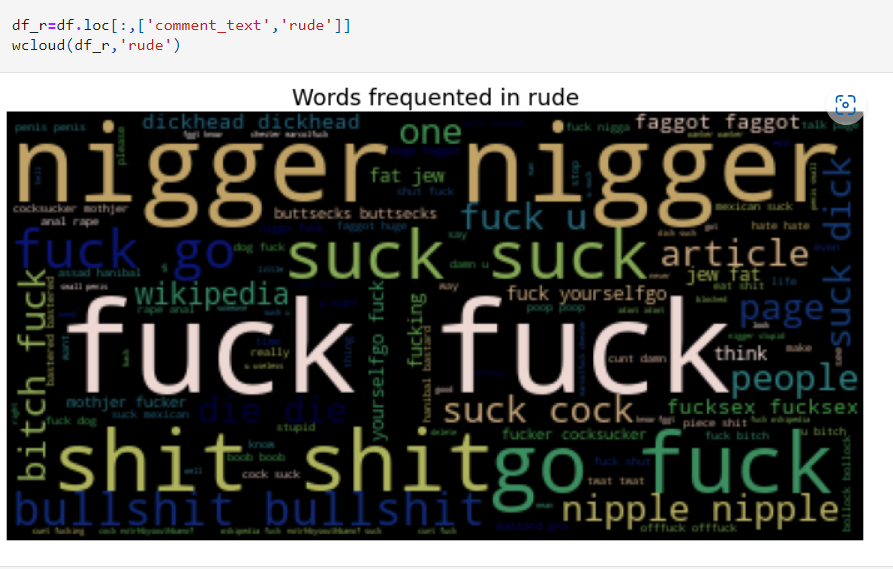
I have analysed the input output logic with word cloud and I have word clouded the sentenced that as classified as foul language in every category. A tag/word cloud is a novelty visual representation of text data, typically used to depict keyword metadata on websites, or to visualize free from text. It’s an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

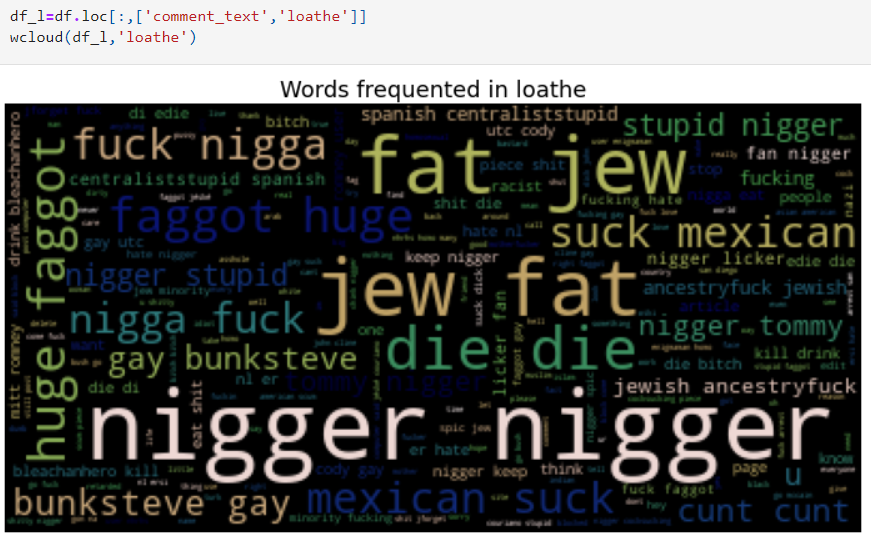
***Code:-***

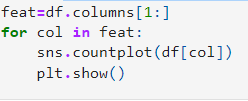
******

***Output:-***

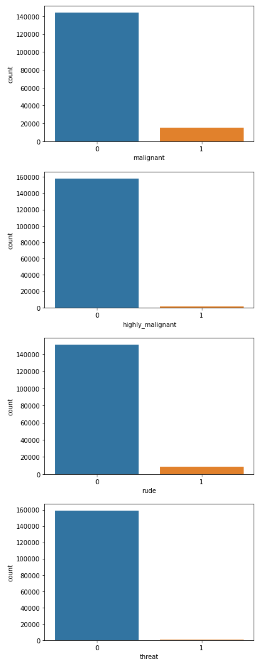
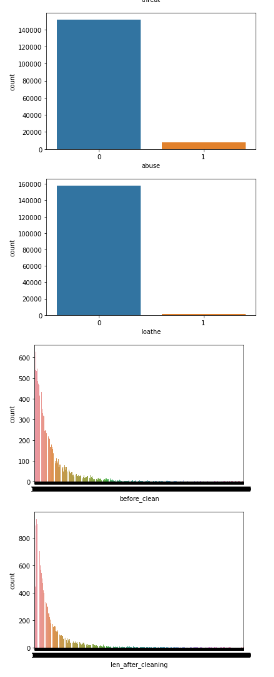
******

******

****Plot all features using countplot:-**

****code:-**

***output:-***

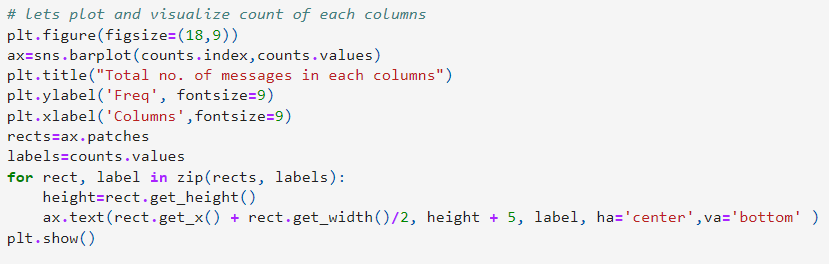
******

**Observation:-**

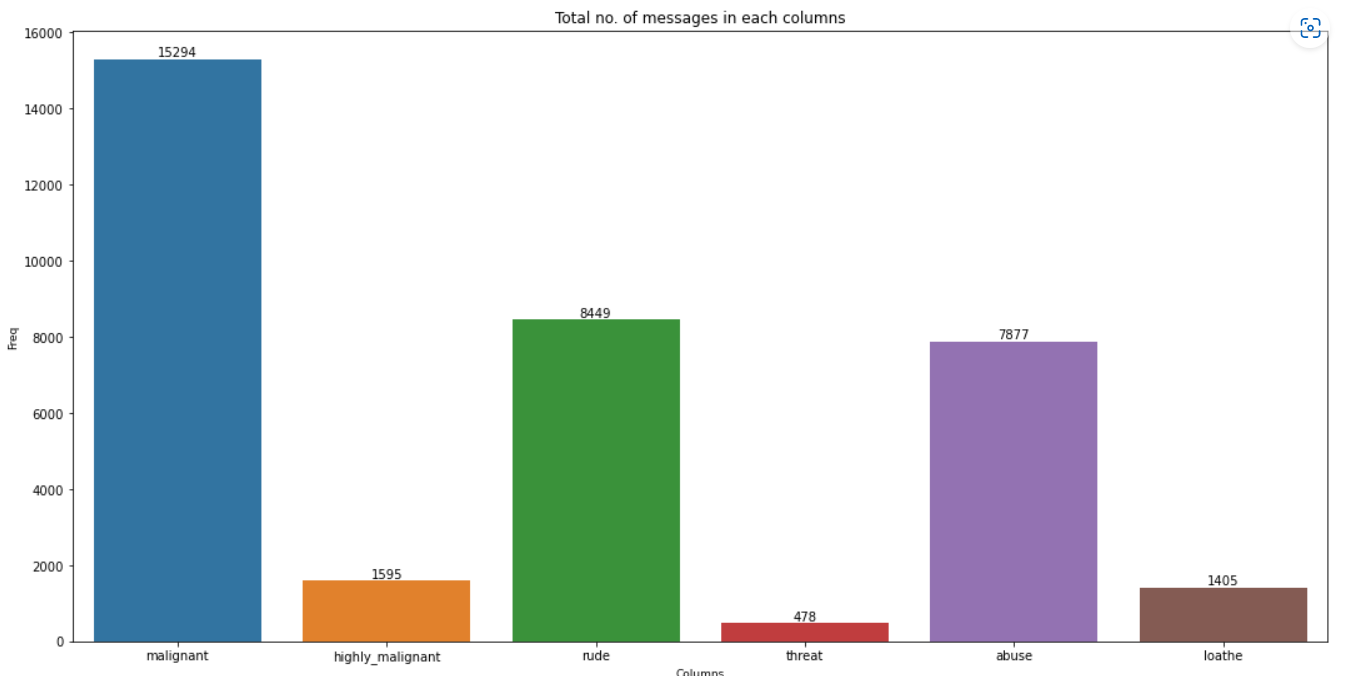
* Here in the first graph of malignant we can clearly observe that most of the messages are not malignant.
* In the second image we can clearly observe that there are very less highly malignant messages.
* Same in third picture there are few rude comments in the dataset.
* In 4th we can clearly see that there are very few cases/almost negligible of threat comments
* In 5th image we can clearly see that there are some messages with abusive language.
* While in the sixth image we can clearly see that there are very few cases of loathe messages.
* In 7th image we can see the no. of words in each rows
* In 8th image we can see the cleaned no. of remaining words in each row.

**plot and visualize count of each columns:-**

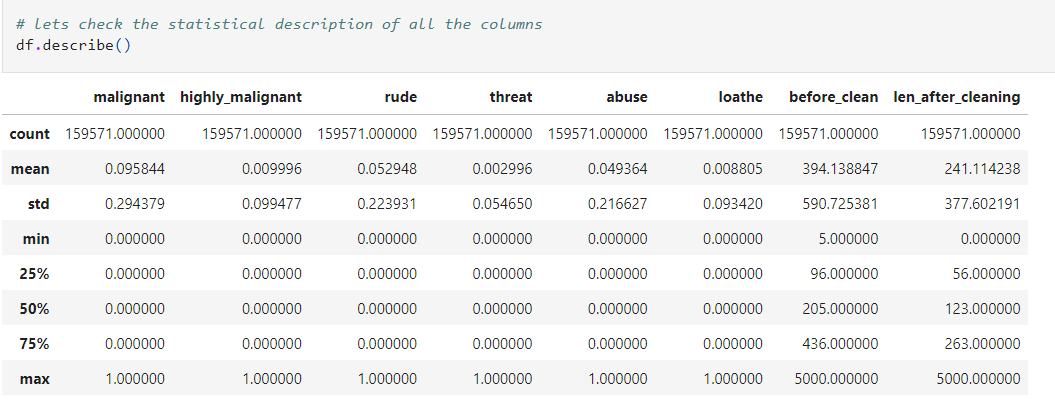
**code:-**

****

**Output:-**

****

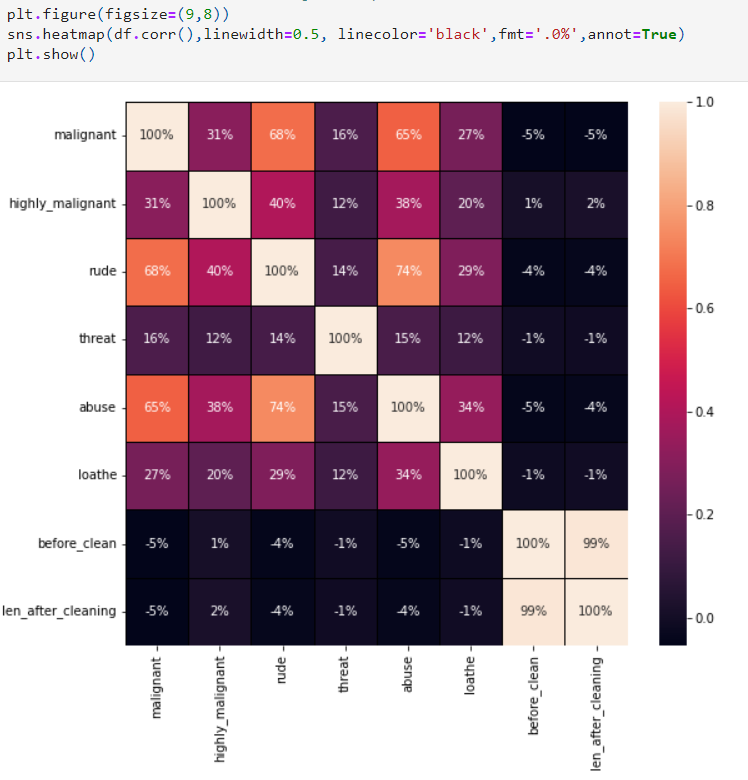
**Describing Dataset:-**

****

***Observation:-***

* Here we can see that only 2 values are present in all the columns i.e. 0 and 1.
* Low score of standard devaiation tells us that the data is not spreaded.
* there is difference in mean and median which tells us that some skewness is present.
* very low difference in 75% and max shows that there are no outliers present in the dataset.

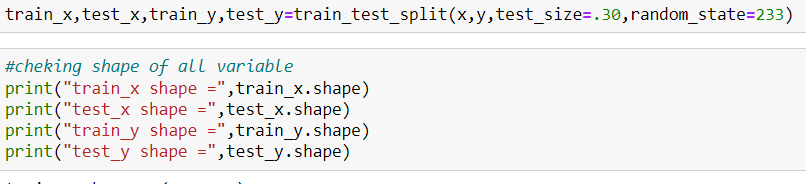
**Correlation Checking:-**

****

**Model Building:-**

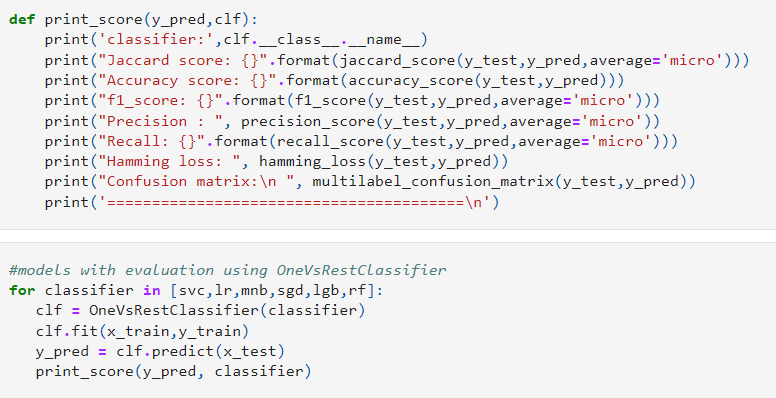
### ***Train\_Test\_Split:-***

The [Train\_Test\_Split](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.



### ***Model Selection:-***





***Run and selecte Best models:-***

classifier: LinearSVC

Jaccard score: 0.8478403520284093

Accuracy score: 0.9176554144385026

f1\_score: 0.9176554144385026

Precision : 0.9176554144385026

Recall: 0.9176554144385026

Hamming loss: 0.08234458556149733

Confusion matrix:

[[[ 2850 2018]

[ 257 42747]]

[[45344 620]

[ 1573 335]]

[[46511 317]

[ 907 137]]

[[45861 748]

[ 657 606]]

[[47134 210]

[ 431 97]]

[[47729 27]

[ 108 8]]

[[47861 2]

[ 9 0]]]

========================================

classifier: LogisticRegression

Jaccard score: 0.8443875093910732

Accuracy score: 0.9156291778074866

f1\_score: 0.9156291778074866

Precision : 0.9156291778074866

Recall: 0.9156291778074866

Hamming loss: 0.08437082219251336

Confusion matrix:

[[[ 2130 2738]

[ 97 42907]]

[[45521 443]

[ 1685 223]]

[[46676 152]

[ 991 53]]

[[45991 618]

[ 661 602]]

[[47261 83]

[ 481 47]]

[[47751 5]

[ 115 1]]

[[47863 0]

[ 9 0]]]

========================================

classifier: MultinomialNB

Jaccard score: 0.8337546924078756

Accuracy score: 0.9093415775401069

f1\_score: 0.9093415775401069

Precision : 0.9093415775401069

Recall: 0.9093415775401069

Hamming loss: 0.09065842245989304

Confusion matrix:

[[[ 1271 3597]

[ 23 42981]]

[[45799 165]

[ 1852 56]]

[[46823 5]

[ 1040 4]]

[[46038 571]

[ 774 489]]

[[47342 2]

[ 526 2]]

[[47756 0]

[ 116 0]]

[[47863 0]

[ 9 0]]]

========================================

classifier: SGDClassifier

Jaccard score: 0.8368858277535829

Accuracy score: 0.9112007018716578

f1\_score: 0.9112007018716578

Precision : 0.9112007018716578

Recall: 0.9112007018716578

Hamming loss: 0.08879929812834225

Confusion matrix:

[[[ 1407 3461]

[ 10 42994]]

[[45906 58]

[ 1899 9]]

[[46731 97]

[ 1001 43]]

[[46075 534]

[ 739 524]]

[[47268 76]

[ 484 44]]

[[47731 25]

[ 109 7]]

[[47863 0]

[ 9 0]]]

========================================

classifier: LGBMClassifier

Jaccard score: 0.8471630042636931

Accuracy score: 0.9172585227272727

f1\_score: 0.9172585227272727

Precision : 0.9172585227272727

Recall: 0.9172585227272727

Hamming loss: 0.08274147727272728

Confusion matrix:

[[[ 2485 2383]

[ 161 42843]]

[[45585 379]

[ 1750 158]]

[[46623 205]

[ 935 109]]

[[45822 787]

[ 585 678]]

[[47207 137]

[ 428 100]]

[[47694 62]

[ 93 23]]

[[47855 8]

[ 9 0]]]

========================================

classifier: RandomForestClassifier

Jaccard score: 0.8460589233379608

Accuracy score: 0.9166109625668449

f1\_score: 0.9166109625668449

Precision : 0.9166109625668449

Recall: 0.9166109625668449

Hamming loss: 0.08338903743315508

Confusion matrix:

[[[ 2429 2439]

[ 195 42809]]

[[45646 318]

[ 1746 162]]

[[46562 266]

[ 886 158]]

[[45775 834]

[ 599 664]]

[[47217 127]

[ 444 84]]

[[47748 8]

[ 113 3]]

[[47863 0]

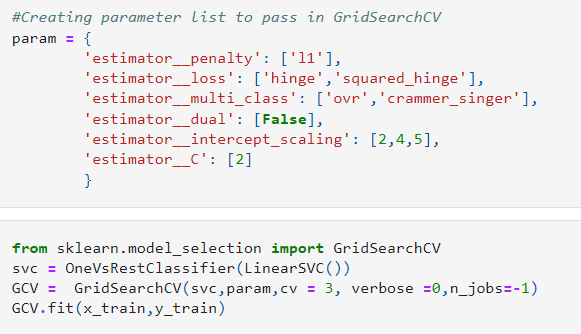
[ 9 0]]]

========================================

we get best accuracy score from **LinearSVC**

#### **Hyperparameter Tuning:-**

***Code:-***



***Output:-***

GridSearchCV(cv=3, estimator=OneVsRestClassifier(estimator=LinearSVC()),

n\_jobs=-1,

param\_grid={'estimator\_\_C': [2], 'estimator\_\_dual': [False],

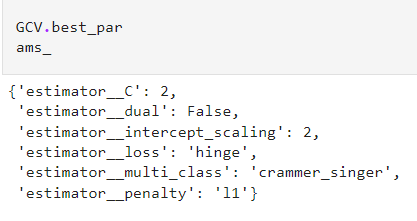
'estimator\_\_intercept\_scaling': [2, 4, 5],

'estimator\_\_loss': ['hinge', 'squared\_hinge'],

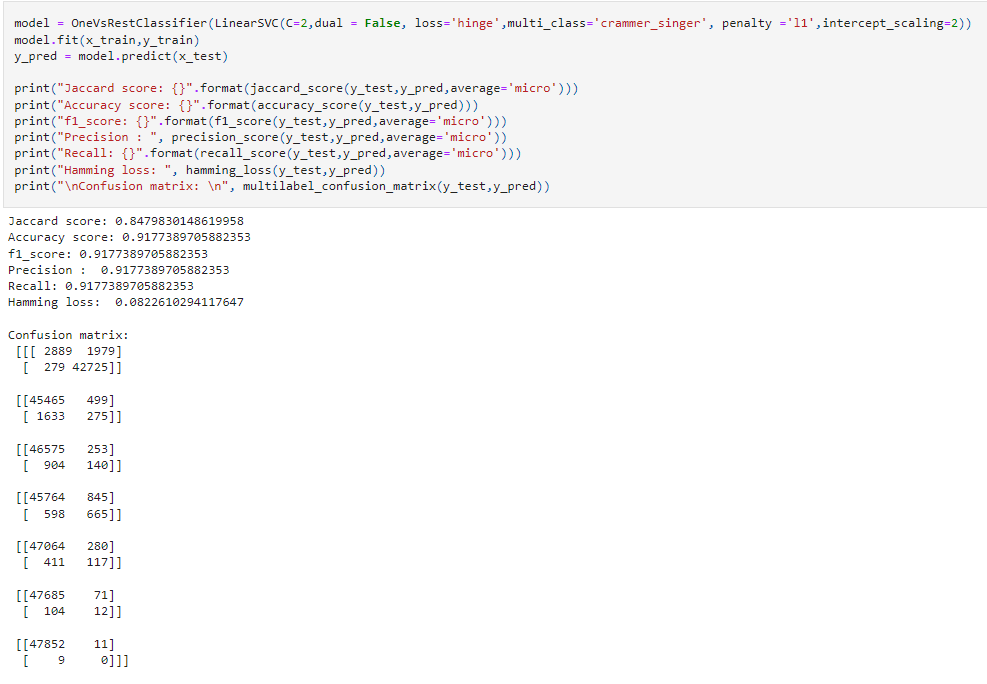
'estimator\_\_multi\_class': ['ovr', 'crammer\_singer'],

'estimator\_\_penalty': ['l1']})

***Checcking Best Parameter:-***

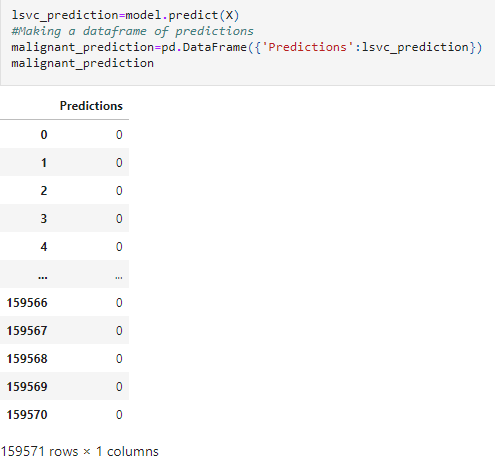
******

**Creat Final Model:-**

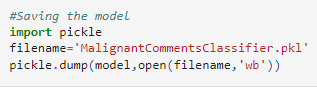


Here we have successfully improved accuracy score from 91.76 to 91.77%.

* ***SHOW PREDICTED RESULTS:-***

******

**Saving Model:-**

****

**Interpretation of the Results:-**

* LinearSVM and Random Forest models perform best.
* At first we get only 91.76% accuracy from LinearSVM But after parameter tuning we get 91.77% accuracy.There is no defference in accuracy score after parameter tuning.
* Using hyper parameter tunning we can improve our model accuracy, But here in this model the accuracy did not increased.
* It is always advised to all of us that atleast we need to use 5 Algorithm in order to figure out which one is performing best among them and we choose that one and we send that for hyper parameter tuning to know that best parameter .

***CONCLUSION***

For any of machine learning project my suggestion is first you have to understand the problem on ground level .if you don’t allow yourself to work with diligence .if you don’ t work harder anything that you are doing or will do , not only in case of machine learning but also in life cycle would be futile. Maybe, my endeavour assist you when ever you will get stuck

* For future improvements, following step we thought to took-
  + Replacing model with a latest/different model
  + Using other robust datasets
  + More focus on NLP properties
* It would seem that better performance might be achieved if multiple learners were combined.

**Reference:-**

I have also used few external resources that helped me to complete this project successfully.Below are the external resources that were used to fulfill my project.

* [**https://www.google.co.in/**](https://www.google.co.in/)
* [**https://github.com/**](https://github.com/)
* [**https://www.kaggle.com/**](https://www.kaggle.com/)
* [**https://www.youtube.com/index**](https://www.youtube.com/index)
* [**https://careerkarma.com/blog/nlp-projects/**](https://careerkarma.com/blog/nlp-projects/)

