NAME OF THE PROJECT

Malignant-Comments-Classifier

SUBMITTED BY:

Wesley Sirra

ACKNOWLEDGMENT

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo.

Chap 1. 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.

Data Set Description:

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

The data set includes:

Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not. Highly Malignant: It denotes comments that are highly malignant and hurtful. Rude: It denotes comments that are very rude and offensive. Threat: It contains indication of the comments that are giving any threat to someone. Abuse: It is for comments that are abusive in nature. Loathe: It describes the comments which are hateful and loathing in nature. ID: It includes unique Ids associated with each comment text given. Comment text: This column contains the comments extracted from various social media platforms.

Analytical Problem Framing

1. Mathematical / Analytical Modelling of the Problem

Whenever we employ any ML algorithm, statistical models or feature pre-processing in background lot of mathematical framework work. In this project we have done lot of data pre-processing & ML model building. In this section we dive into mathematical background of some of these algorithms.

1. Logistic Regression

The response variable, label, is a binary variable (whether the loan was repaid or not). Therefore, the logistic regression is a suitable technique to use because it is

developed to predict a binary dependent variable as a function of the predictor variables. The logit, in this model, is the likelihood ratio that the dependent variable, non-defaulter, is one (1) as opposed to zero (0), defaulter. The probability, P, of credit default is given by;

$$\ln\left[\frac{P(Y)}{1-P(Y)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Where;

$ln\left[\frac{P(Y)}{1-P(Y)}\right]$ is the log (odds) of credit default

Y is the dichotomous outcome which represents credit default (whether the loan was repaid or not) $X_1, X_2, \ldots, X_K$ are the predictor variables which are as educational level, number of dependents, type of loan, adequacy of the loan facility, duration for repayment of loan, number of years in business, cost of capital and period within the year the loan was advanced to the client $\beta_o, \beta_1, \beta_2, \ldots, \beta_k$ are the regression (model) coefficients

2. Data Sources and their formats

The data set comes from my internship company – Fliprobo technologies in excel format.

```
# Importing dataset CSV file using pandas
df= pd.read_csv('Data file.csv')
```

```
print('No. of Rows :',df.shape[0])
print('No. of Columns :',df.shape[1])
pd.set_option('display.max_columns',None) # # This will enable us to see truncated columns
df.head()
```

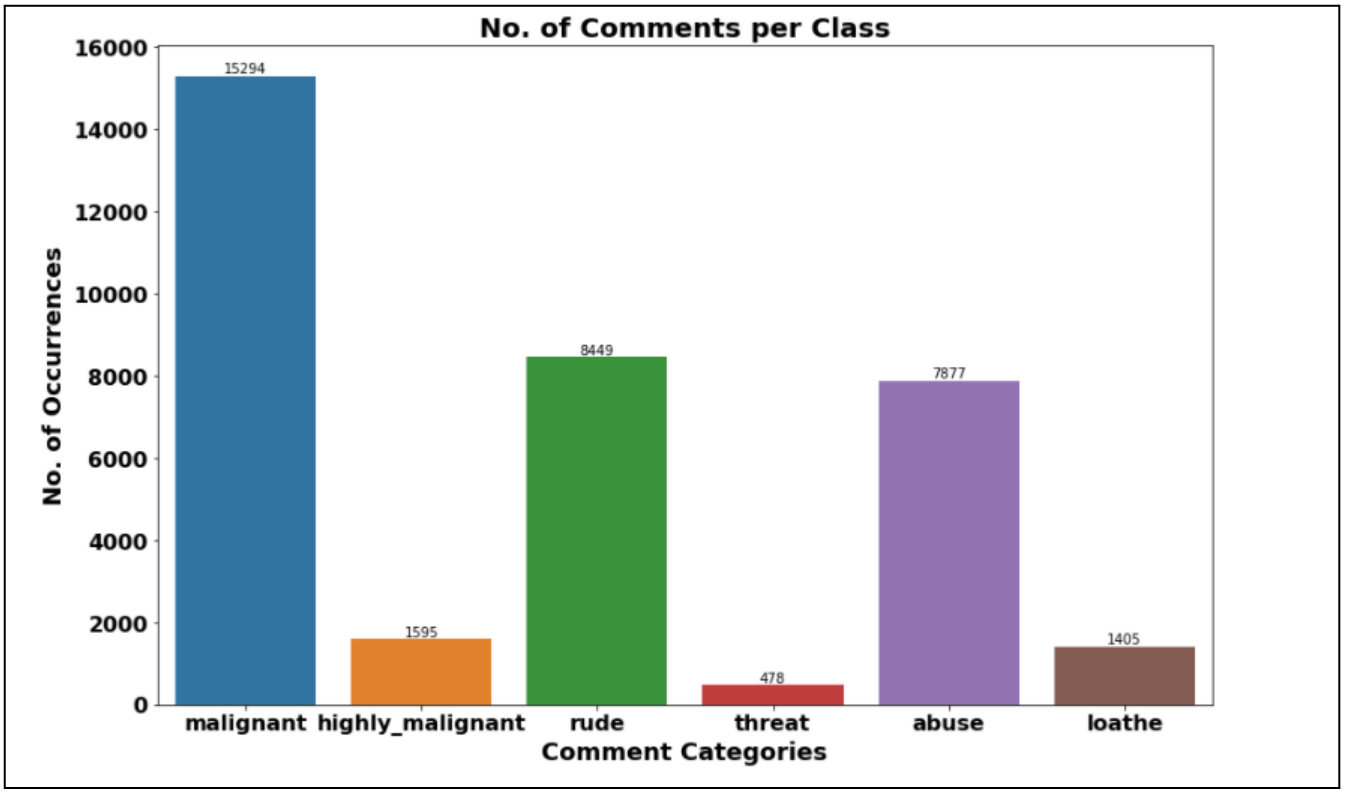
```
No. of Rows : 209593
No. of Columns : 37
```

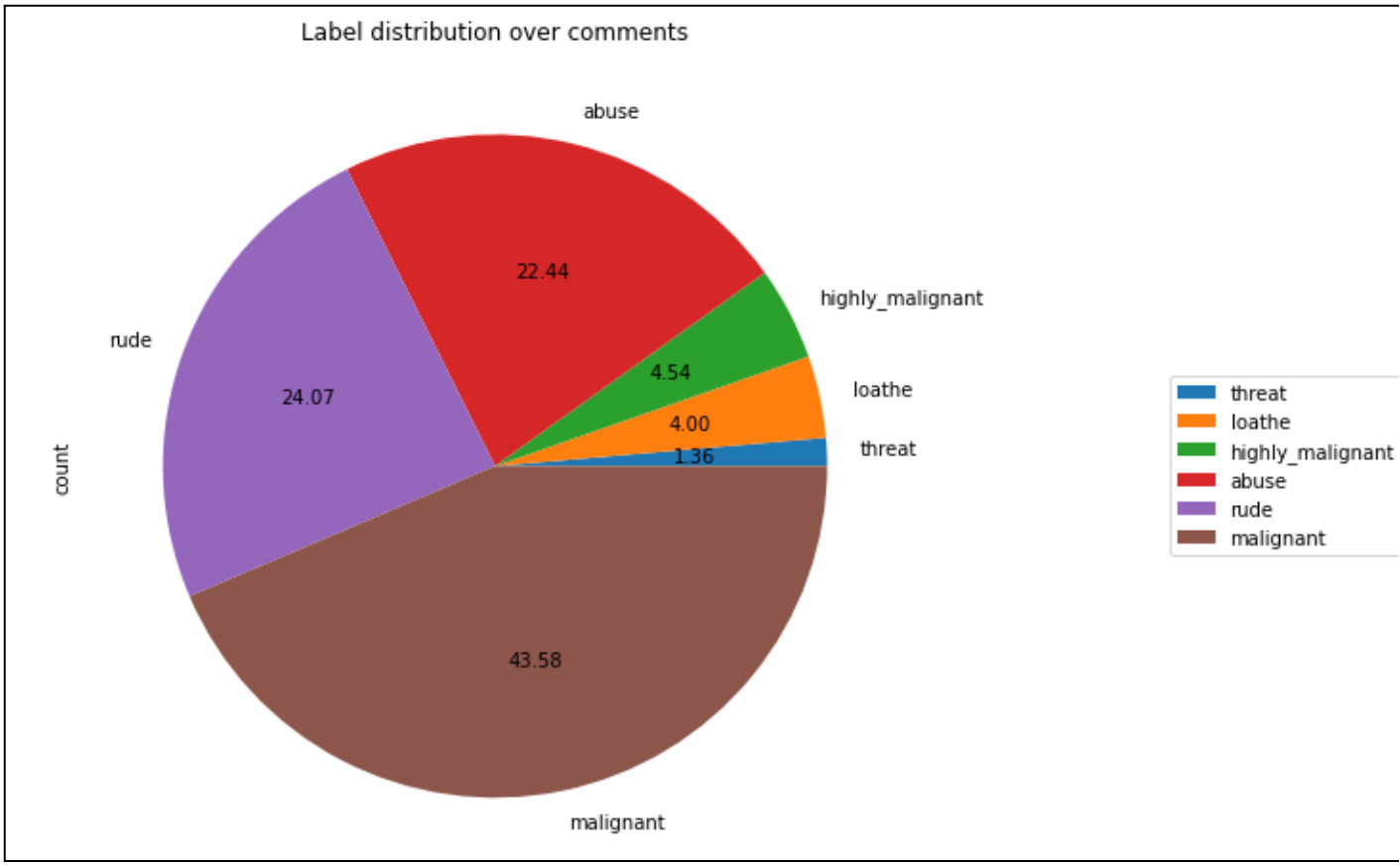
Malignant Commentes Classifier - Multi Label Classification

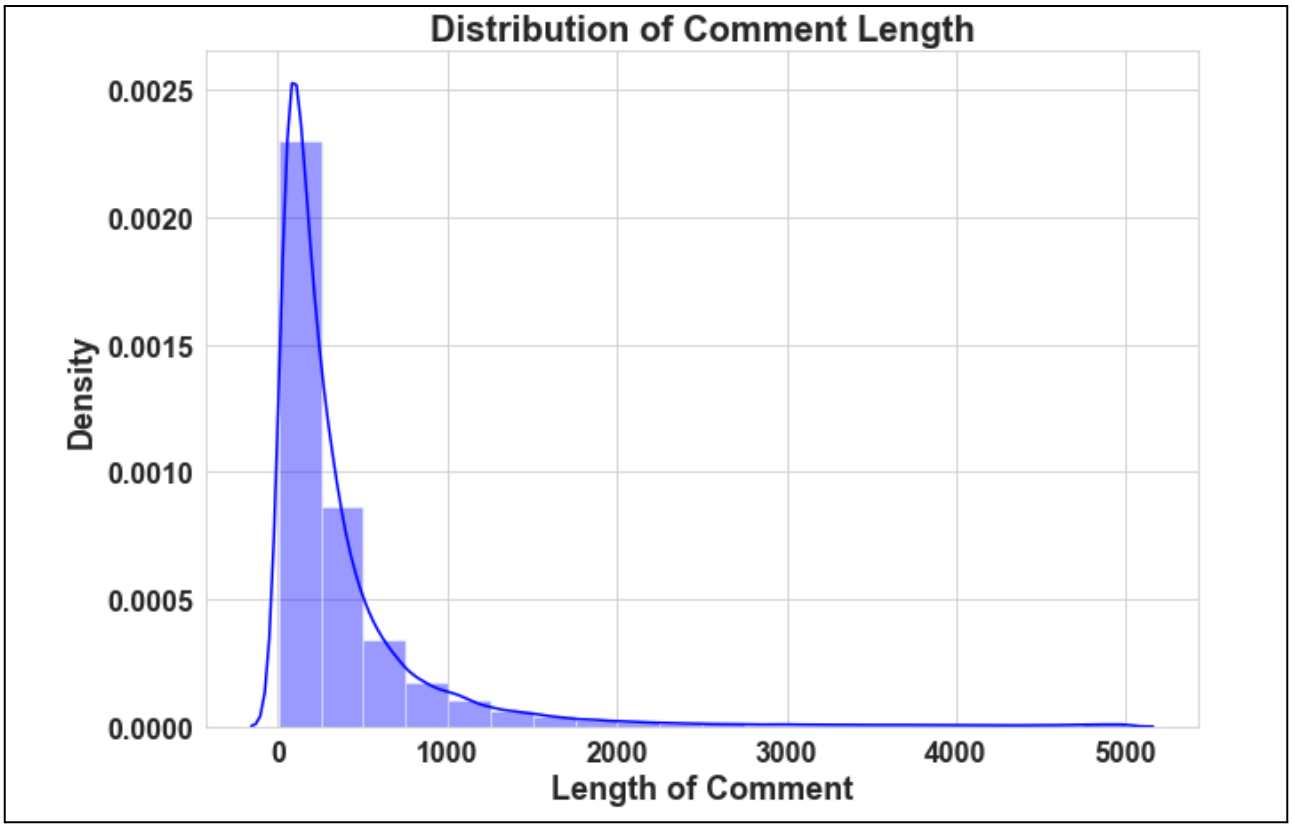
- There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading 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.
- 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.
-

Exploration of Target Variable Ratings :-







Data Pre Processing

- Convert the text to lowercase
- Remove the punctuations, digits and special characters

- Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
- Remove the stop words
- Stemming and Lemmatising
- Applying Text Vectorization to convert text into numeric

Multi-Label Classification Techniques

- **One Vs Rest**
- **Binary Relevance**
- **Classifier Chains**
- **Label Powerset**
- **Adapted Algorithm**

Word Cloud:-

Word Cloud is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.

The more commonly the term appears within the text being analysed, the larger the word appears in the image generated.

The enlarged texts are the greatest number of words used there and small texts are the smaller number of words use.


WORDS TAGGED AS MALIGNANT
fucking
whitey
edits
hey
absurd
around
cocksucker
taliban
come
antisemmitian
group
piss
talk
think
bye
mother
exclusive
tosser
work
comming
destr
gay
wp
look
back
good
filthy
hair
pubic
shit
annoyed
great
deleted
Length
posting
keep
g
meow
going
libe
wwe
tiger
archangel
dry
mischievous
previous
fuck
delete
sit
dtype
object
Name
u
stuff
listen
eating
shark
comment_text
conversation


WORDS TAGGED AS HIGHLY MALIGNANT
asshole
motherfucking
pathetic
shalom
dtype
u
stupid
blocking
hi
de
antonio
mercado
Length
kill
lol
co
know
walter
stuff
shit
around
stop
crow
cocksucker
object
semite
work
fuck
crap
moron
frankly
cunt
woman
peace
son
better
deleting
im
fuckhead
jimbo
feel
quite
wi
Name
good
stink
piss
back
fucker
stinking
gay
cc
comment_text
wale
rootmyass
bitch
fucking
dirty
leave
never
mother
piece
filthy


WORDS TAGGED AS RUDE
fucking
piece
shit
white
stuff
absurd
sedits
gay
around
antisemmitian
fucked
naked
fuckeeed
mother
great
libe
object
lolwat
got
conversation
stupid
fuck
dry
u
fat
tiger
work
put
meow
previous
eating
going
thought
shark
cant
posting
faggot
peace
obese
gpiss
deleting
keep
asshole
stop
dtype
sit
cocksucker
drink
archangel
filthy
comment_text
Length
Name
deleted

Visualization & Data Wrangling Library used

```
#Importing warning library to avoid any warnings
import pandas as pd # for data wrangling purpose
import numpy as np # Basic computation library
import seaborn as sns # For Visualization
import matplotlib.pyplot as plt # ploting package
%matplotlib inline
import warnings # Filtering warnings
warnings.filterwarnings('ignore')
```

Text Mining Library used

```
#Importing required libraries
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from wordcloud import WordCloud
```

Machine Learning Model Building Library used

```
#Importing Machine Learning Model Library
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
from skmultilearn.problem_transform import BinaryRelevance
from sklearn.svm import SVC, LinearSVC
from sklearn.multiclass import OneVsRestClassifier
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.metrics import confusion_matrix,classification_report,accuracy_score
from sklearn.metrics import roc_auc_score, roc_curve, auc
from sklearn.metrics import hamming_loss, log_loss
```

The different classification algorithm used in this project to build ML model are as below:

- ❖ **Random Forest classifier**
- ❖ **Support Vector Classifier**
- ❖ **Logistics Regression**
- ❖ **AdaBoost Classifier**

Machine Learning Evaluation Matrix

- **Support Vector Classifier gives maximum <u>Accuracy Score: 91.1508 %</u> and <u>Hamming Loss: 2.0953%</u> than the other classification models.**
- **Hyper parameter Tuning is perform over this best model using best param shown below :**

```
Out[69]: {'estimator__loss': 'hinge',
          'estimator__multi_class': 'ovr',
          'estimator__penalty': 'l2',
          'estimator__random_state': 42}
```

Final ML Model

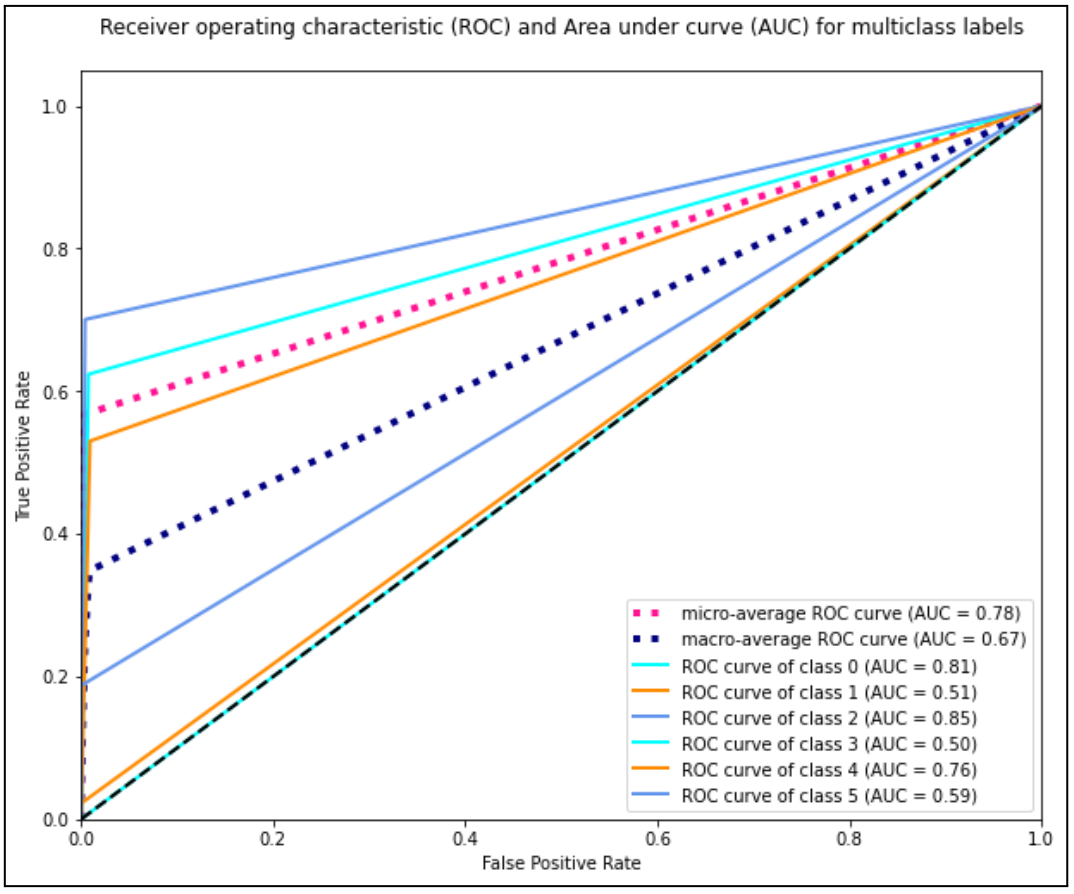
Final Model

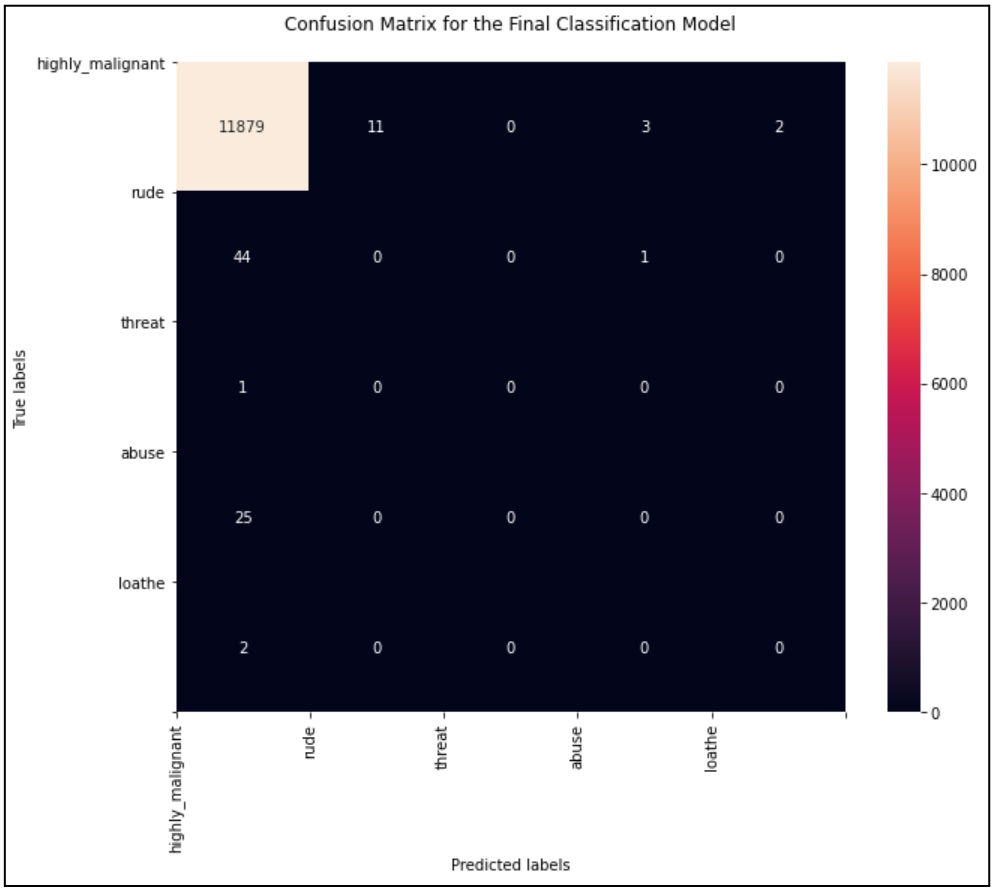
```
Final_Model = OneVsRestClassifier(LinearSVC(loss='hinge',
            multi_class='ovr', penalty='l2', random_state=42))

Classifier = Final_Model.fit(x_train, y_train)
fmod_pred = Final_Model.predict(x_test)
fmod_acc = (accuracy_score(y_test, fmod_pred))*100
print("Accuracy score for the Best Model is:", fmod_acc)
h_loss = hamming_loss(y_test,fmod_pred)*100
print("Hamming loss for the Best Model is:", h_loss)
```

```
Accuracy score for the Best Model is: 91.26002673796792
Hamming loss for the Best Model is: 2.0819407308377897
```

Final Model is giving us Accuracy score of 91.26% which is slightly improved compare to earlier Accuracy score of 91.15%.

AOC-ROC Curve & Confusion Matrix





Machine Learning Evaluation Matrix

Algorithm	Accuracy Score	Recall (Micro)
Logistics Regression	0.9123	0.89
Random Forest Classifier (RFC)	0.9074	0.56
Support Vector Classifier	0.9115	0.56
Ada Boost Classifier	0.9057	0.50

Learning Outcomes of the Study in respect of Data Science

1. First time I handle such huge dataset.
2. First time any project I worked on ever need such data clean operation. I paid attention realistic & unrealistic data, considering it corrective measure taken as per need. This was beyond normal missing value imputation for me.
3. As data was huge require high computational capacity, it made me switch to Google Colab for running model and for hyperparameter Tuning. I Hyper Tuned Final model with Google Colab GPU.
4. I run Hyper parameter tuning 2-3 times with serval parameter. It was taking lot of times so at end I reduce Hyperparameter search parameter and still it was taken 6-7 hr for finding best parameter.

Limitations of this work and Scope for Future Work Limited
computational resources put limitation on optimization through hyper parameter tuning. Accuracy of model can increase with hyperparameter tuning with several different parameter. Here we use only two parameters for tuning.