NAME OF THE PROJECT

MALIGNANT COMMENTS CLASSIFICATION

Submitted by:

Prathamesh Nayak

ACKNOWLEDGMENT

I would like to express my special gratitude to the “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analysis skills. And I want to express my huge gratitude to Ms. Khushboo Garg (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo. Last but not least my parents have been my backbone in every step of my life.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towards data science, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)
4. Data Science Projects with Python Second Edition by Packt
5. Hands on Machine learning with scikit learn and tensor flow by Aurelien Geron
6. B. Smith, J. Leimkuhler, R. Darrow, and Samuels, “Yield management at American airlines, “Interfaces, vol. 22, pp. 8–31, 1992
7. William Groves, Maria Gini, “An agent for optimizing airline ticket purchasing”, in international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013)
8. Chen, Y., Cao, J., Feng, S., Tan, Y., 2015. An ensemble learning based approach for building airfare forecast service. In: 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 964-969.

9. Yeamduan Narangajavana, Fernando.J. Garrigos-Simon, Javier Sanchez García, Santiago Forgas-Coll, "Prices, prices and prices: A study in the airline sector", Tourism Manage., 41 (2014), pp. 28-42
10. Bo An, Haipeng Chen, Noseong Park, V.S. Subrahmanian MAP: Frequency-Based Maximization of Airline Profits based on an Ensemble Forecasting Approach Proceedings of the 22nd ACM 3 Flight Price Prediction Using ML Techniques SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), ACM, New York, NY, USA (2016), pp. 421-430
11. R. Ren, Y. Yang, and S. Yuan, "Prediction of airline ticket price," University of Stanford, 2014.
12. T. Janssen, T. Dijkstra, S. Abbas, and A. C. van Riel, "A linear quantile mixed regression model for prediction of airline ticket prices," Radboud University, 2014.
13. K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," in the 25th IEEE European signal processing conference, 2017, pp. 1036– 1039.
14. C. Koopmans and R. Lieshout, "Airline cost changes: To what extent are they passed through to the passenger?" Journal of Air Transport Management, vol. 53, pp. 1–11, 2016.
15. G. Francis, A. Fidato, and I. Humphreys, "Airport–airline interaction: the impact of low-cost carriers on two European airports," Journal of Air Transport Management, vol. 9, no. 4, pp. 267–273, 2003.
16. Boruah A., Baruah K., Das B., Das M.J., Gohain N.B. (2019) "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter," https://doi.org/10.1007/978-981-13-0224-4_18
17. G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm," doi:10.1016/j.jpdc.2016.09.001
18. T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.
19. Stacey Mumbower, Laurie A. Garrow, Matthew J. Higgins "Estimating flight-level price elasticities using online airline data: a

first step toward integrating pricing, demand, and revenue optimization", Transportation Res. Part A: Policy Practice, 66 (2014), pp. 196-212

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 behavior.

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 an online hate and abuse comment classifier which can be used to classify hate and offensive comments

so that it can be controlled and restricted from spreading hatred and cyberbullying.

Motivation for the Problem Undertaken

The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skills in solving a real time problem has been the primary motivation.

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 a good amount of data exploration and derive some interesting features using the comments text column available.

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 behavior.

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.

Analytical Problem Framing

Multi –Label Classification Problem

Difference between multi-class classification & multi-label classification is that in multi-class problems the classes are mutually exclusive, whereas for multi-label problems each label represents a different classification task, but the tasks are somehow related.

For example, multi-class classification makes the assumption that each sample is assigned to one and only one label: a fruit can be either an apple or a pear but not both at the same time. Whereas, an instance of multi-label classification can be that a text might be about any religion, politics, finance or education at the same time or none of these.

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 include '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.**

Data Source

```
# Importing dataset excel file using pandas.
df=pd.read_csv('train.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 : 159571
No. of Columns : 8
```

There are 8 features in the dataset . The data types of different features are as shown below

```
# Sorting out columns for datatypes
df.columns.to_series().groupby(df.dtypes).groups
```

```
{int64: ['malignant', 'highly_malignant', 'rude', 'threat', 'abuse', 'loathe'], object: ['id', 'comment_text']}
```

Data Pre-processing

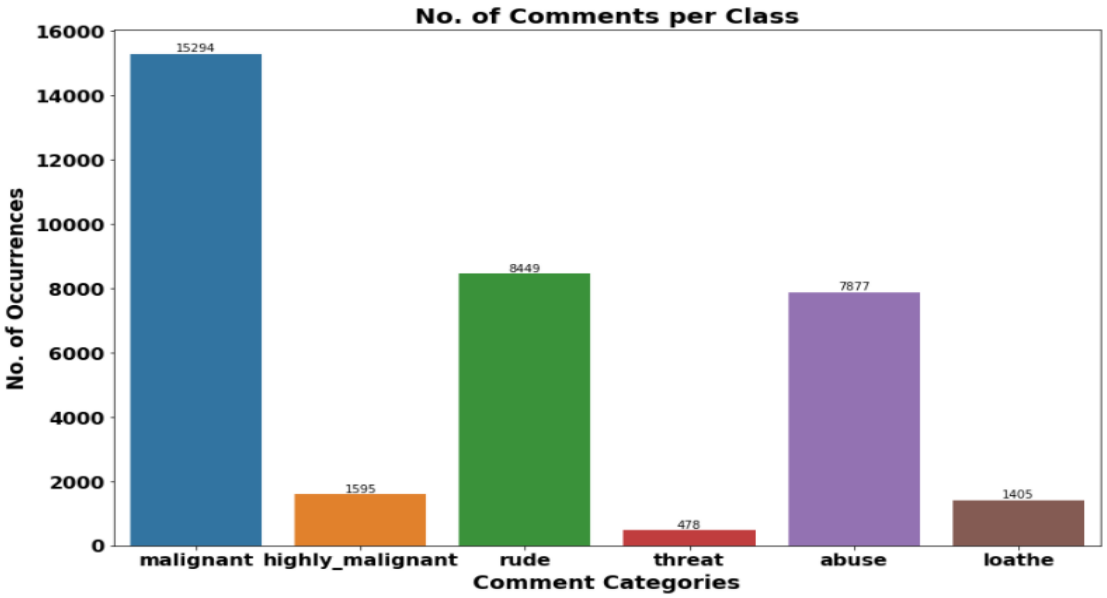
The dataset is large and it may contain some data errors. In order to reach clean, error free data some data cleaning & data pre-processing performed data.

• Data Integrity check – No missing values or duplicate entries present in dataset.

- **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 Lemmatization
- Applying Text Vectorization to convert text into numeric

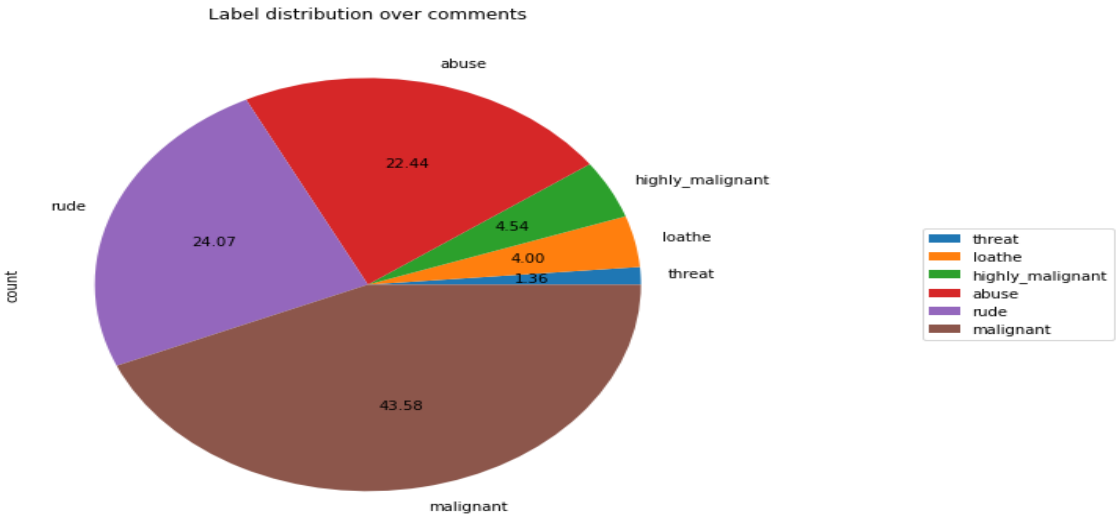
Exploratory Data Analysis



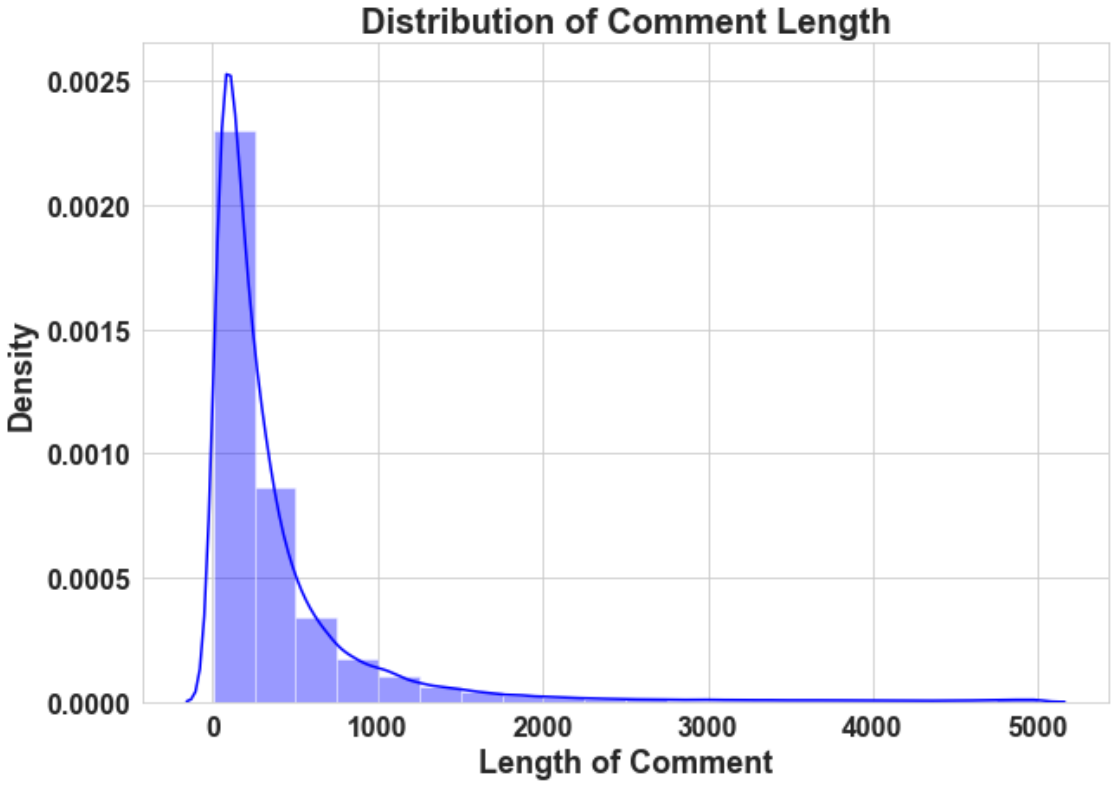
Out of total Negative comments the maximum negative comments come with Malignant in nature followed by rude categories.

Around 90% comments are Good/Neutral in nature while the rest 10% comments are Negative in nature.

Very few comments come with threatening nature.



Out of total negative comments around 43.58% are malignant in nature followed by 24.07% are rude comments.



Above is a plot showing the comment length frequency. As noticed, most of the comments are short with only a few comments longer than 1000 words.

Majority of the comments are of length 500, where maximum length is 5000 and minimum length is 5. Median length being 250.

Multi-Label Classification Techniques

One Vs Rest
Binary Relevance
Classifier Chains
Label Powerset
Adapted Algorithm

Word Cloud for getting word sense

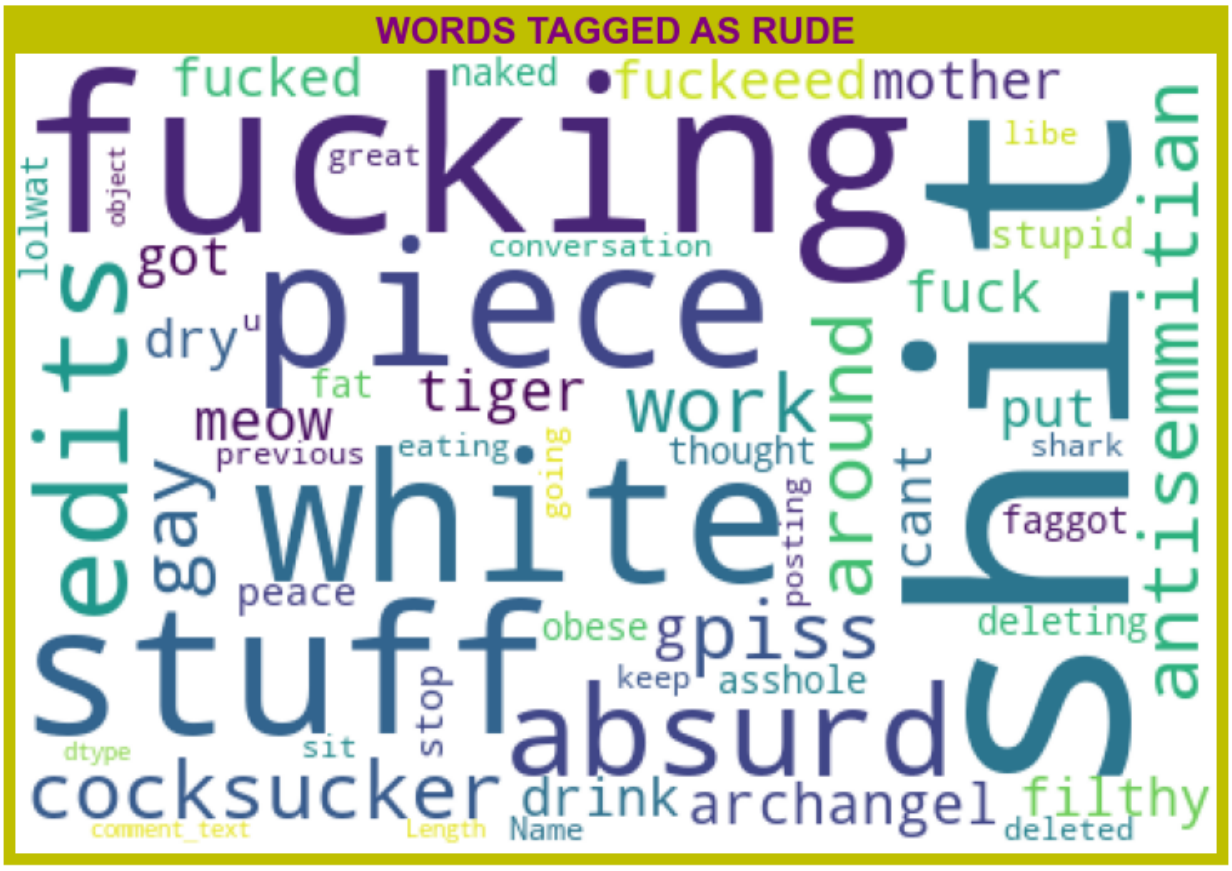
Word Cloud is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.

The more commonly the term appears within the text being analyzed, 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 used.

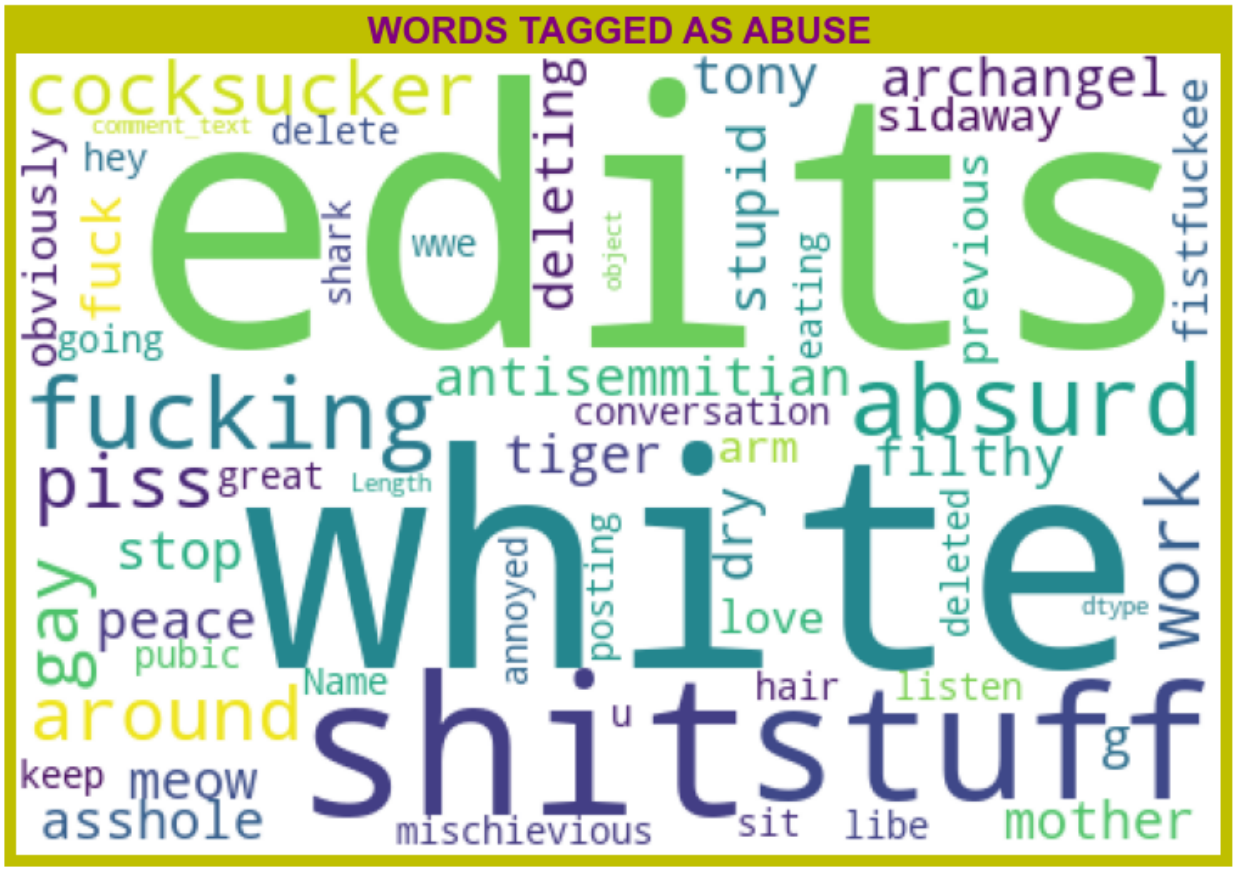
From word cloud of malignant comments, it is clear that it mostly consists of words like edits, hey, white, fucking, gay, cocksucker, work, think, Taliban etc

From word cloud of Highly malignant comments, it is clear that it mostly consists of words like fuck, stupid, fucking, bitch, crow, shit, cocksucker etc

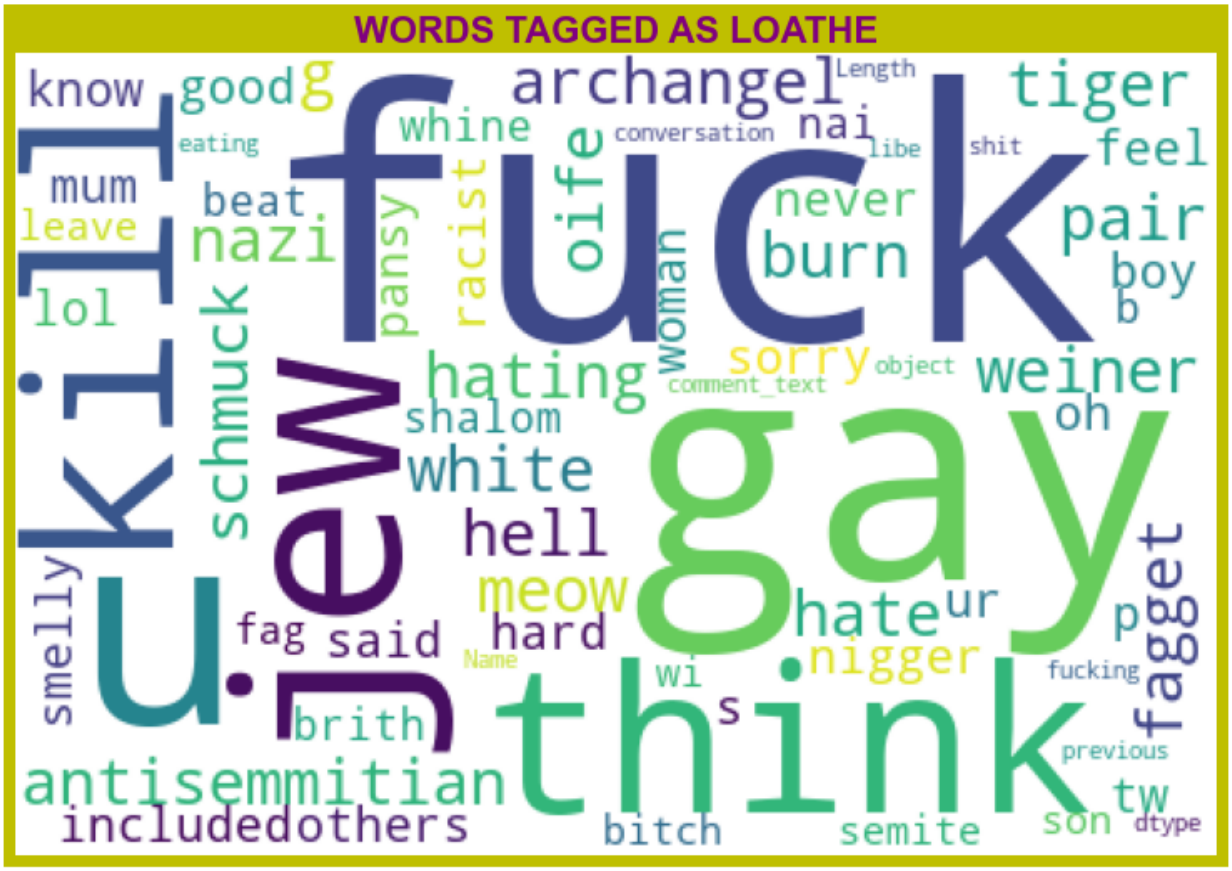


From word cloud of Rude comments, it is clear that it mostly consists of words like fucking, shit, white, piece, edits, stuff, absurd etc.

From word cloud of Threat comments, it is clear that it mostly consists of words like fuck, suck, Bitch, die, stupid, etc.



From word cloud of Abuse comments, it is clear that it mostly consists of words like edits, white, shit, stuff, fuck, piss, fucking etc.



From word cloud of Loathe comments, it is clear that it mostly consists of words like fuck, gay, kill, think, jew, u etc.

Machine Learning Model Building

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 Accuracy Score: 91.1508 % and Hamming Loss: 2.0953% 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}
```

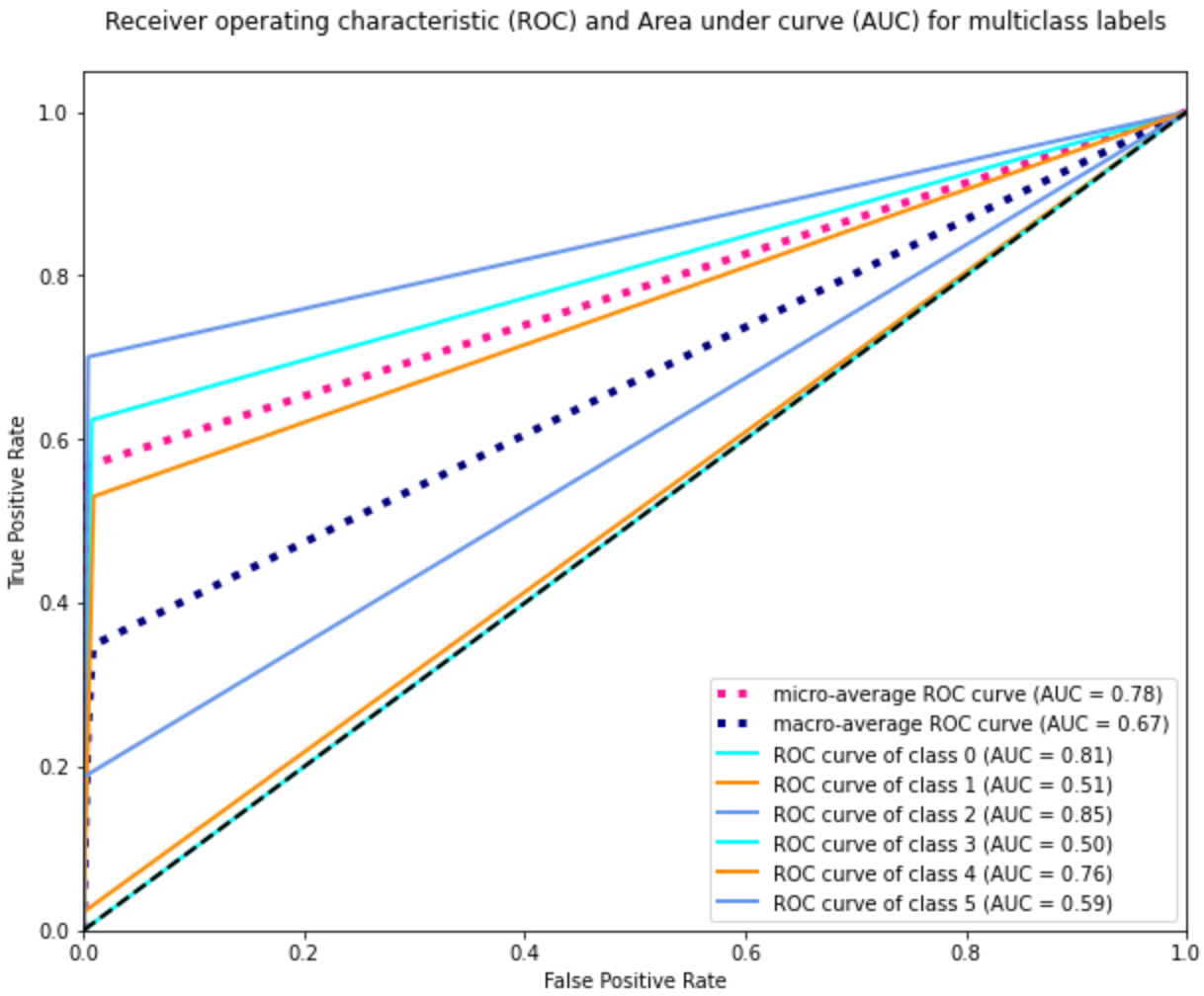
Final Model

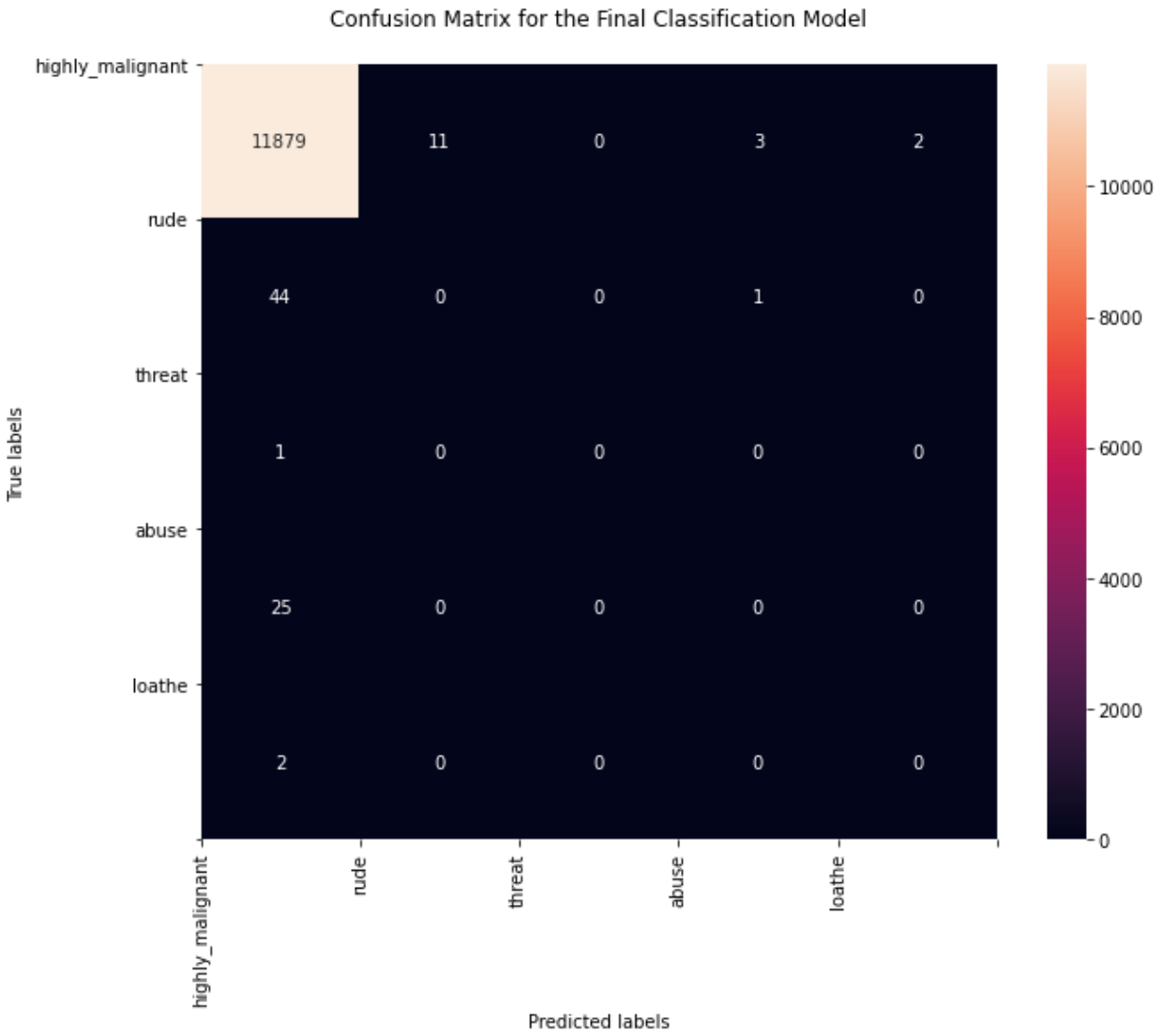
```
Final_Model = OneVsRestClassifier(LinearSVC(loss='hinge',
            multi_class='ovr', penalty='l2', random_state=42))

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 an Accuracy score of 91.26% which is slightly improved compared to earlier Accuracy score of 91.15%.

AOC-ROC Curve & Confusion Matrix


Receiver operating characteristic (ROC) and Area under curve (AUC) for multiclass labels
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
micro-average ROC curve (AUC = 0.78)
macro-average ROC curve (AUC = 0.67)
ROC curve of class 0 (AUC = 0.81)
ROC curve of class 1 (AUC = 0.51)
ROC curve of class 2 (AUC = 0.85)
ROC curve of class 3 (AUC = 0.50)
ROC curve of class 4 (AUC = 0.76)
ROC curve of class 5 (AUC = 0.59)



CONCLUSION

Linear Support Vector Classifier performs better with Accuracy Score: 91.15077857956704 % and Hamming Loss: 2.0952019242942144 % than the other classification models.

Final Model (Hyperparameter Tuning) is giving us an Accuracy score of 91.26% which is slightly improved compared to earlier Accuracy score of 91.15%.

SVM classifier is the fastest algorithm compared to others.

Limitations of this work and Scope for Future Work

The Maximum feature used while vectorization is 2000. Employing more features in vectorization leads to a more accurate model which I am not able to employ due computational resources.

Data is imbalanced in nature but due to computational limitations we have not employed balancing techniques here.

Deep learning CNN, ANN can be employed to create more accurate models.