

**NAME OF THE PROJECT**

Malignant-Comments-Classifier

**Submitted by:**

**AKANKSHA MISHRA**

**FLIPROBO SME:**

**MS. KHUSHBOO GaRG**

## 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. 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 who have been my backbone in every step of my life.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towardsdatascience, 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. Stackoverflow.com to resolve some project related queries.
7. Predicting Credit Default among Micro Borrowers in Ghana Kwame Simpe Ofori, Eli Fianu
8. Predicting Microfinance Credit Default: A Study of Nsoatreman Rural Bank, Ghana Ernest Yeboah Boateng
9. A Machine Learning Approach for Micro-Credit Scoring Apostolos Ampountolas

And also thank you for many other persons who has helped me directly or indirectly to complete the project.

# 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.

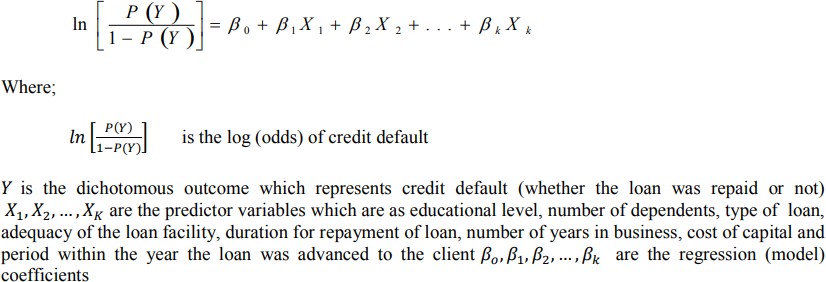
# Analytical Problem Framing

## 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.

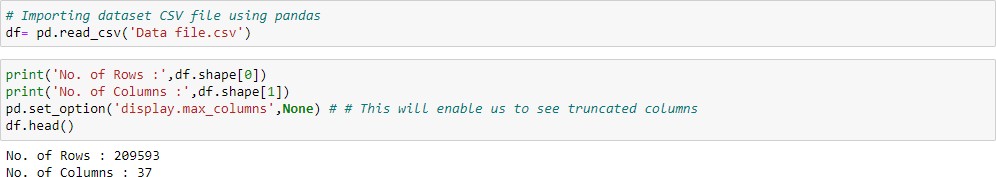
1. Logistic Regression

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;



## Data Sources and their formats

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



## Malignant Commentes Classifier - Multi Label Classification

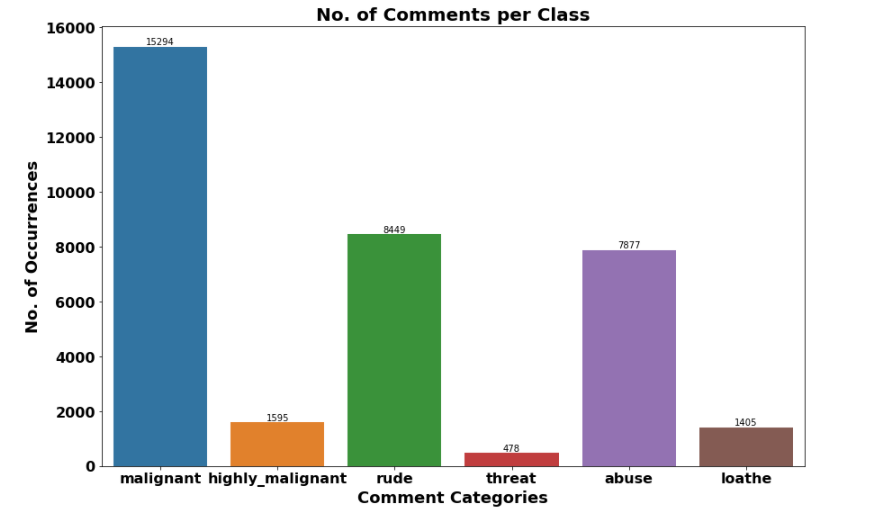
## There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading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

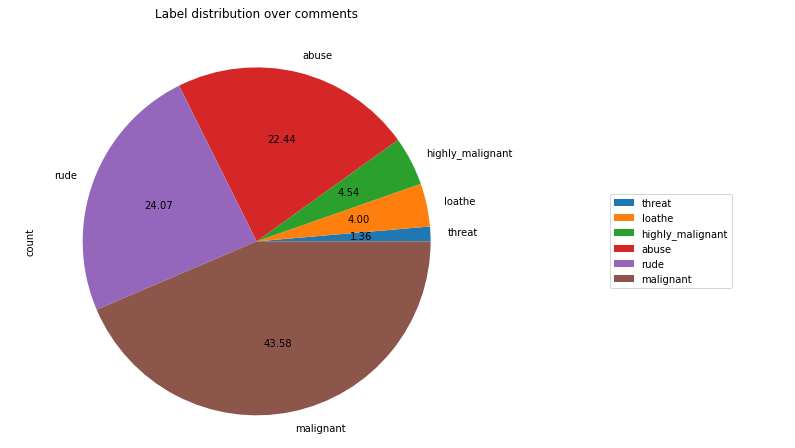
## Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

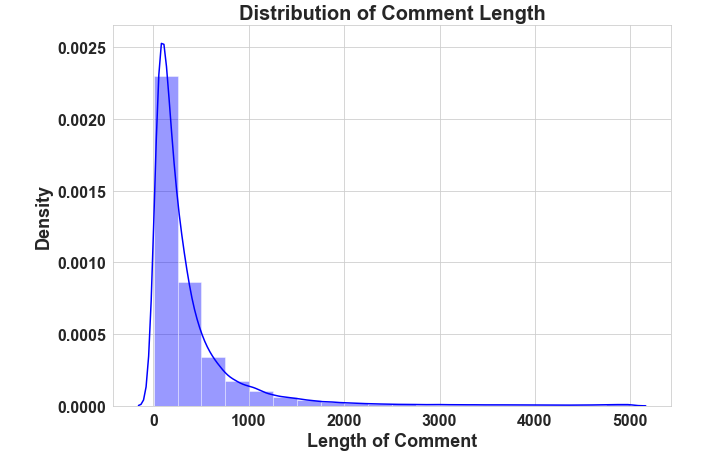
## There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading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

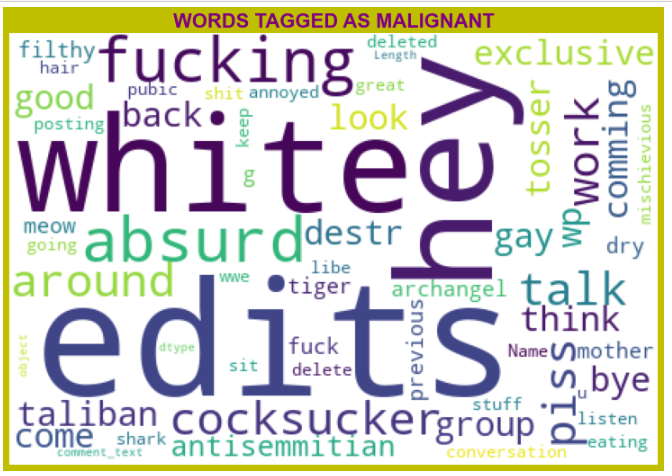
* Convert the text to lowercase
* Remove the punctuations, digits and special characters
* Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
* Remove the stop words
* Stemming and Lemmatising
* Applying Text Vectorization to convert text into numeric

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.

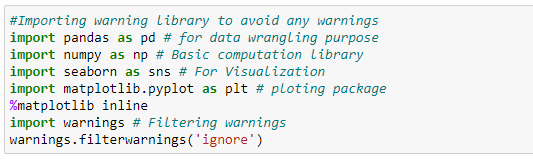




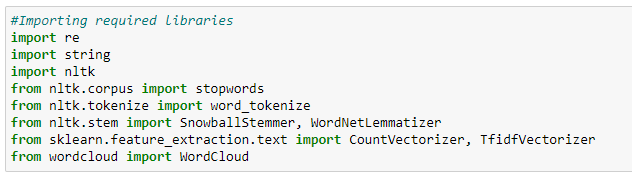


****

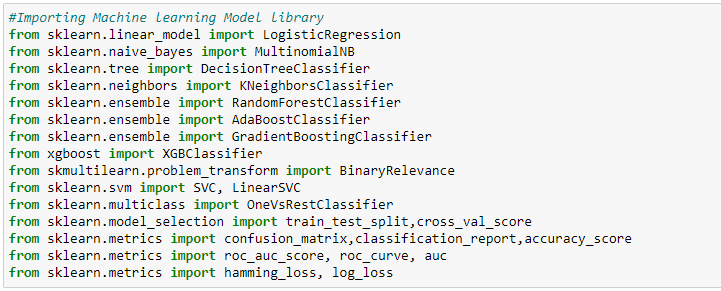
**Visualization & Data Wrangling Library used**

****

**Text Mining Library used**

****

**Machine Learning Model Building Library used**

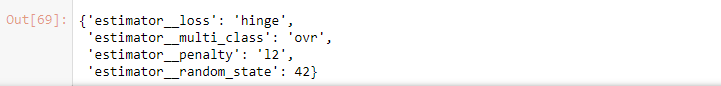
****

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

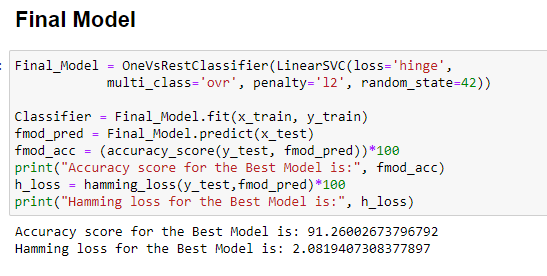
* **Random Forest classifier**
* **Support Vector Classifier**
* **Logistics Regression**
* **AdaBoost Classifier**

**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 :**

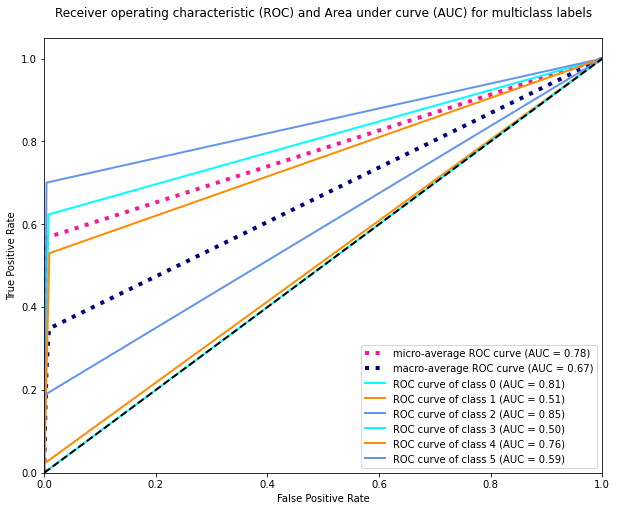
****

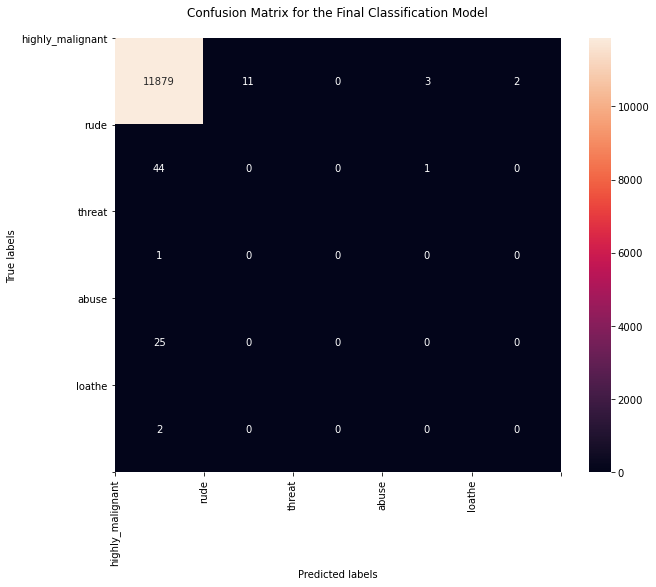
**Final ML Model**

****

**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)** | **Precision**  **(Micro)** | **F1 Score**  **(Micro)** | **Humming Loss** |
| **Logistics Regression** | **0.9123** | **0.89** | **0.94** | **0.61** | **0.02206** |
| **Random Forest Classifier (RFC)** | **0.9074** | **0.56** | **0.79** | **0.66** | **0.02191** |
| **Support Vector Classifier** | **0.9115** | **0.56** | **0.82** | **0.67** | **0.02952** |
| **Ada Boost Classifier** | **0.9057** | **0.50** | **0.80** | **0.61** | **0.9057** |

## 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

1. 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.