

NAME OF THE PROJECT :

MALAIGNANT COMMENTS

CLASSIFICATION.

**Submitted By   
Farhan Baig Balekundri**

# ACKNOWLEDGMENT

Foremost, I would like to express my sincere gratitude to Data TrainedteamforthecontinuoussupportofmyDataSciencestudyand research, for the patience, motivation, enthusiasm, and immense knowledge. The guidance helped me in all the time of research and writingofthisthesis.Icouldnothaveimaginedhavingabetteradvisor and mentor for my Data sciencestudy.

BesidesDataTrained,IwouldliketothankFlipRoboTeam,fortheir encouragement, insightful internship, and help to understand the study.

# INTRODUCTION

## Business ProblemFraming

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 problemacrossmultipleplatforms,thereisalackofmodelsforonline hatedetection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a

majorthreatononlinesocialmediaplatforms.Socialmediaplatforms are the most prominent grounds for such toxicbehaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influencesarefacingbacklashesfrompeopleandhavetocomeacross hateful and offensive comments. This can take a toll on anyone and affectthemmentallyleadingtodepression,mentalillness,self-hatred and suicidalthoughts.

Internet comments are bastions of hatred and vitriol. While online anonymityhasprovidedanewoutletforaggressionandhatespeech, 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.

## Conceptual Background of the Domain Problem

Online platforms and social media become the place where people sharethethoughtsfreelywithoutanypartialityandovercomingallthe race people share their thoughts and ideas among thecrowd.

Social media is a computer-based technology that facilitates the sharing of ideas, thoughts, and information through the building of virtualnetworksandcommunities.Bydesign,socialmediaisInternet- based and gives users quick electronic communication of content. Content includes personal information, documents, videos, and photos. Users engage with social media via a computer, tablet, or smartphone via web-based software orapplications.

While social media is ubiquitous in America and Europe, Asian countries like India lead the list of social media usage. More than 3.8 billion people use social media.

Inthishugeonlineplatformoranonlinecommunitytherearesome peopleorsomemotivatedmobwilfullybullyotherstomakethemnot to share their thought in rightful way. They bully others in a foul language which among the civilized society is seen as ignominy. And when innocent individuals are being bullied by these mob these individuals are going silent without speaking anything. So, ideally the motive of this disgraceful mob isachieved.

To solve this problem, we are now building a model that identifies allthefoullanguageandfoulwords,usingwhichtheonlineplatforms like social media principally stops these mob using the foul language in an online community or even block them or block them from using this foullanguage.

## Review ofLiterature

### The purpose of the literature review is to:

1. Identify the foul words or foul statements that are beingused.
2. Stopthepeoplefromusingthesefoullanguagesinonlinepublic forum.

To solve this problem, we are now building a model using our machine language technique that identifies all the foul language and foul words, using which the online platforms like social media principally stops these mob using the foul language in an online community or even block them or block them from using this foul language.

I have used 5 different Classification algorithms and shortlisted the best on basis on the metrics of performance and I have chosen one algorithm and build a model in that algorithm.

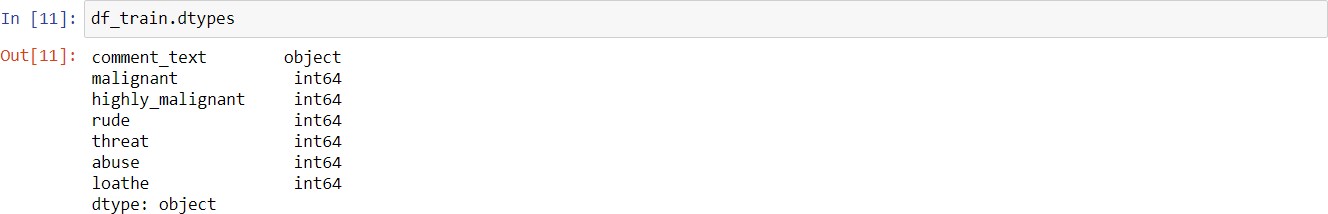
## Motivation for the Problem Undertaken

One of the first lessons we learn as children is that the louder you screamandthebiggerofatantrumyouthrow,youmoreyougetyour way. Part of growing up and maturing into an adult and functioning memberofsocietyislearninghowtouselanguageandreasoningskills to communicate our beliefs and respectfully disagree with others, using evidence and persuasiveness to try and bring them over to our way of thinking. Social media is reverting us back to those animalistic tantrums, schoolyard taunts and unfettered bullying that define youth, creating a dystopia where even renowned academics and dispassionate journalists transform from Dr. Jekyll into raving Mr. Hydes, raising the critical question of whether social media should simply enact a blanket ban on profanity and name calling? Actually, ban should be implemented on these profanities and taking that as a motivation I have started this project to identify the malignant comments in social media or in online publicforms.

# Analytical ProblemFraming

## Mathematical/ Analytical Modeling of the Problem

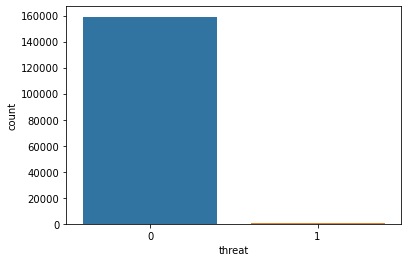
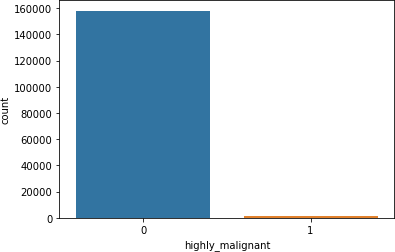
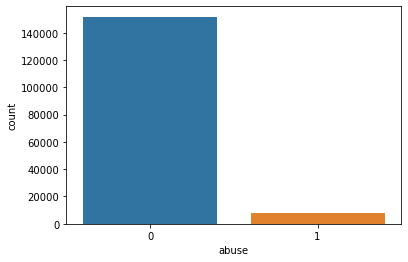
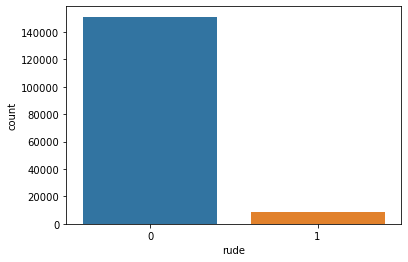
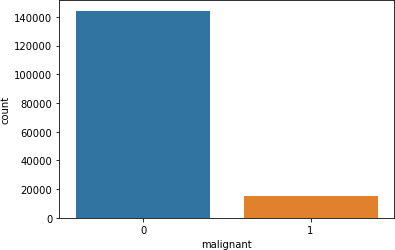
I start analysis on this project in importing the data set and simple play around with the data and identifying the characteristics of each column.



In the first stoke of analysis I understood that there are 8 columns in total in which 6 are numerical column with binary data 0’s and 1’s and ‘id’ data which has all unique values “connect text” have string values.

Since ‘id’ have all unique values, it won’t be helpful in analysis to I

have dropped id column.

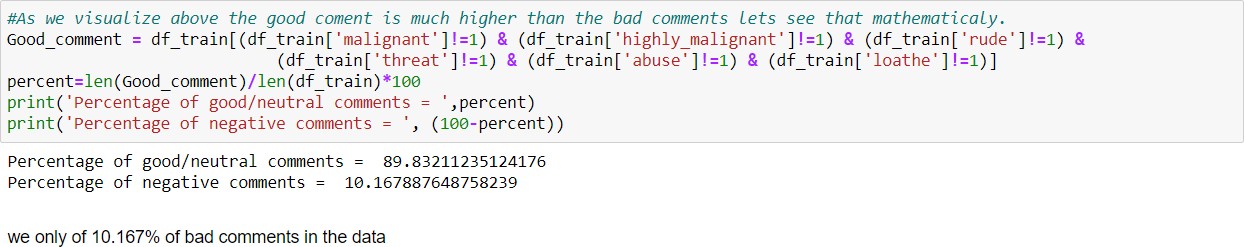




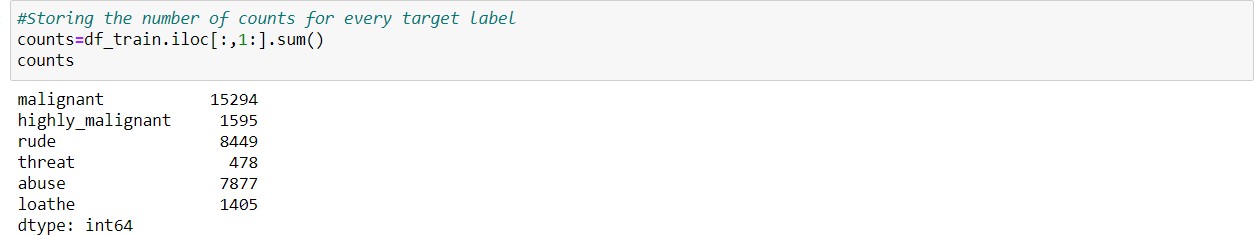
Post dropping ‘id’ I tried to understand data from 'malignant', 'highly\_malignant','rude','threat','abuse','loathecolumnsbyplotting them in countplot.

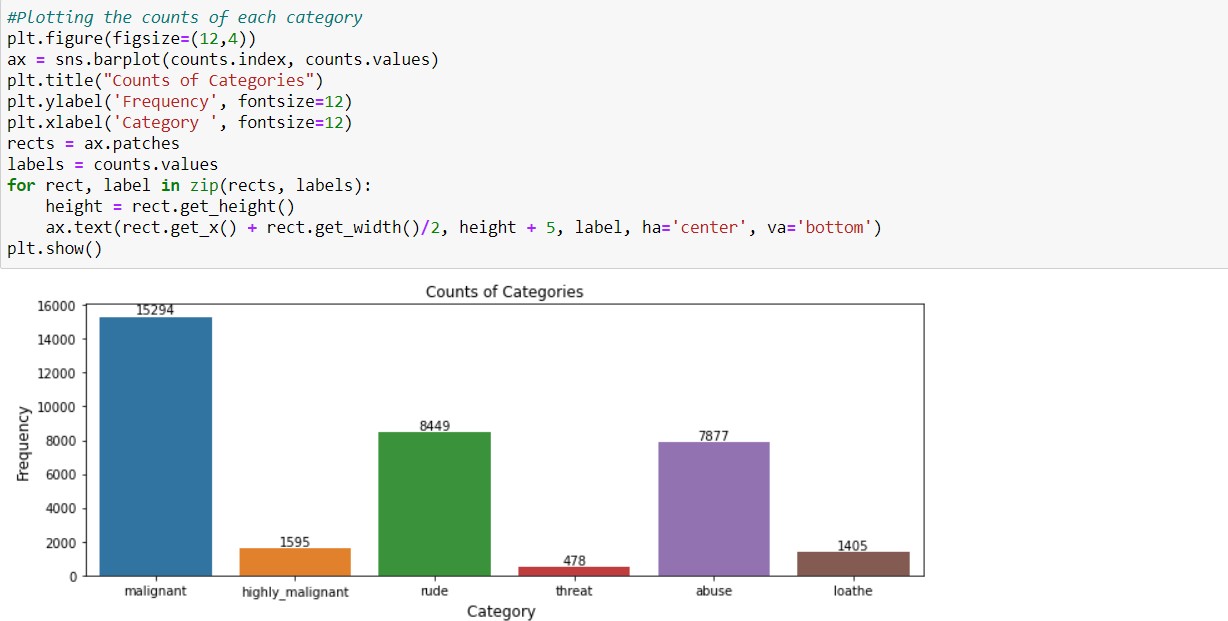
### Key observation:

we can see that there only minimum number of columns in 'malignant', 'highly\_malignant', 'rude', 'threat', 'abuse', 'loathe' and remaining all in 0.



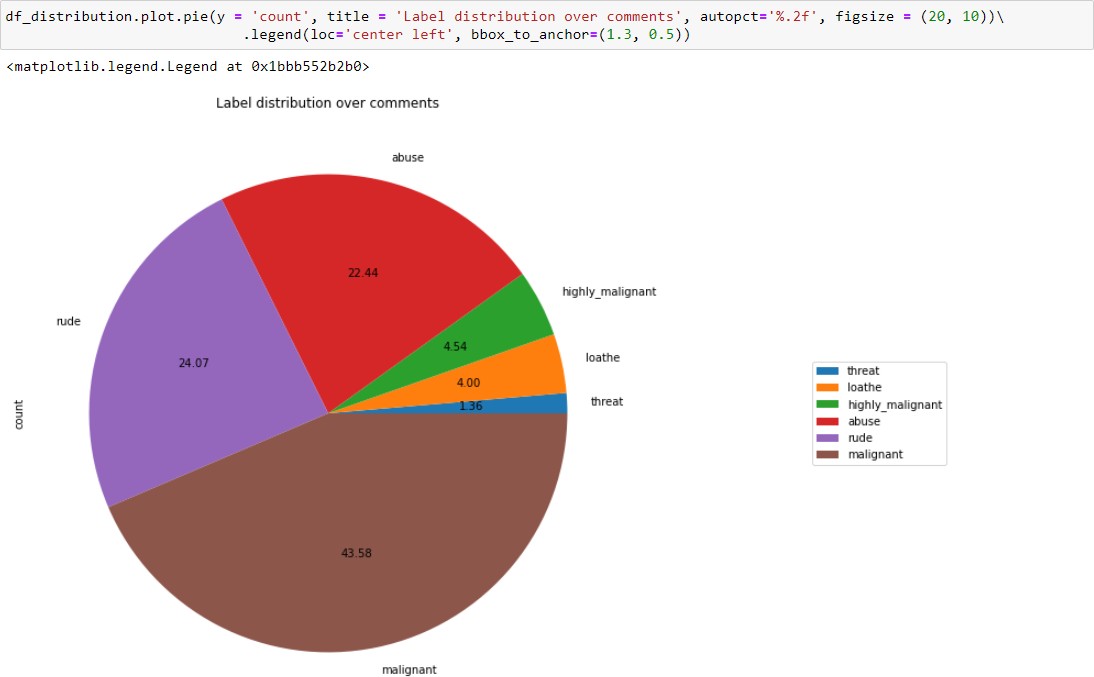
So, only 10% of the data is getting classified as malignant comments data is unbalanced.





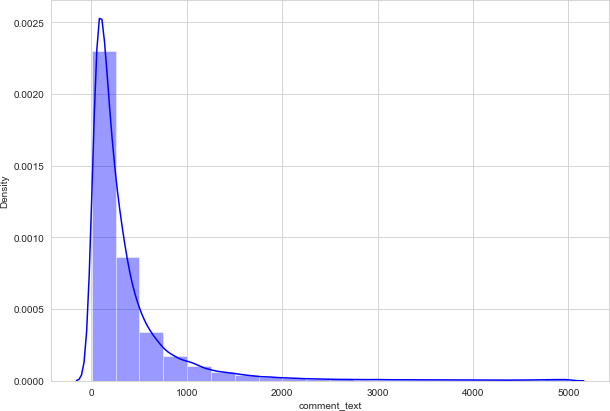
### Key observation:

1. We can see malignant and rude are high categorised sentences in the data.



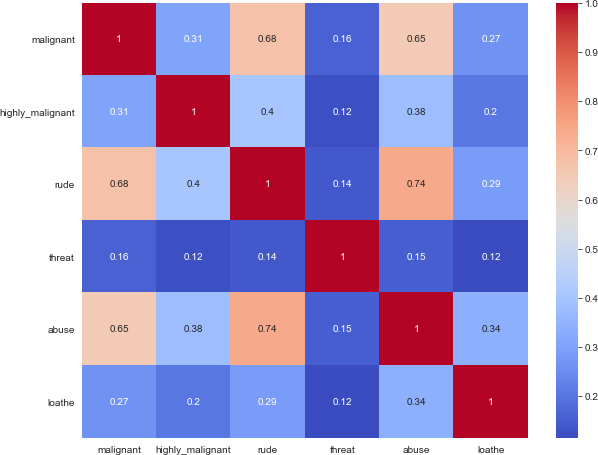
### Key observation:

1. As we see above malignant, and rude sentence are high classified and threat, loathe are least classified.



### Key observation:

1. We can see that few sentences are really long but most of the sentence are small.



### Key Observations:

1. We can see more corelations in the variables, Abuse have more corelation with malignant andrude.
2. Rude has more positive corelation withmalignant
3. we don’t have any negative corelations in thedata.

## Data Sources and their formats

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.

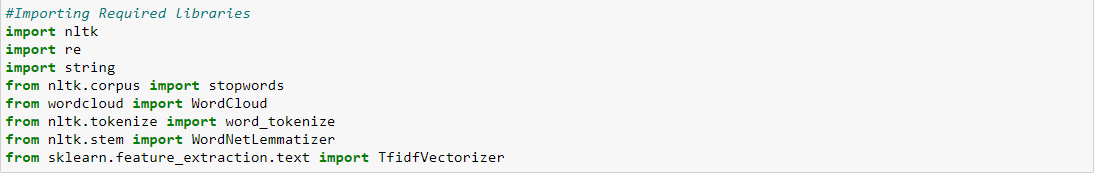
### The data setincludes:

1. Malignant:ItistheLabelcolumn,whichincludesvalues0and1, denoting if the comment is malignant ornot.
2. Highly Malignant: It denotes comments that are highly malignant andhurtful.
3. Rude: It denotes comments that are very rude andoffensive.
4. Threat: It contains indication of the comments that are giving any threat tosomeone.
5. Abuse: It is for comments that are abusive innature.
6. Loathe: It describes the comments which are hateful and loathing innature.
7. ID: It includes unique Ids associated with each comment text given.
8. Comment text: This column contains the comments extracted from various social mediaplatforms.

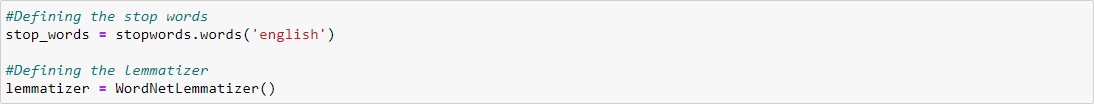
This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes many categories of comments, we can do good amount of data exploration and derive some interesting features using the comments text column available.

## Data Pre-processingDone

I imported all the required libraries for cleansing the data.

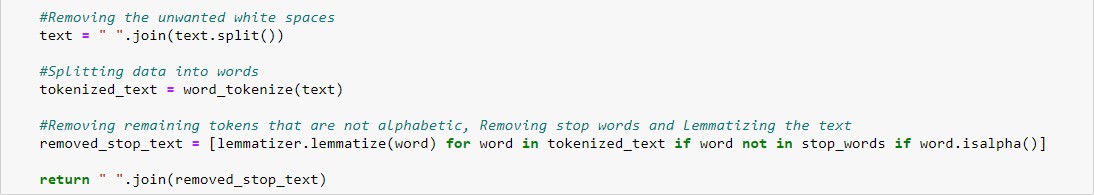


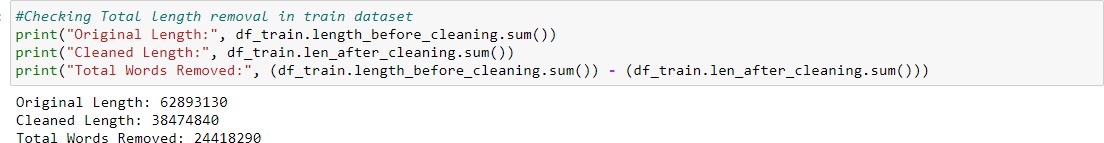
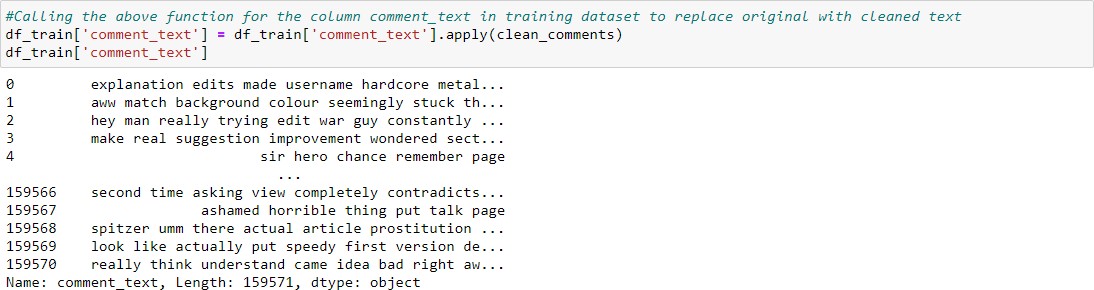
Afterimportingalltherequiredlibraries,Ihavedefinedstopwordsand lemmatize to avariable.



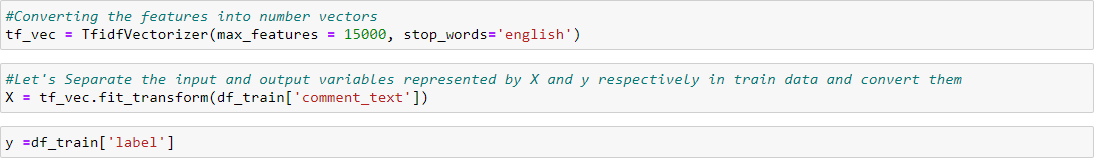
Post which I have defined a function for cleaning the data.





Post on creating a function I have passed my data into the same to clean it.

The total amount of data that is cleansed from the original data is 24418290.Nowthedataiscleansedandreadyfortrainingbutbefore which I converted the data into vectors for the machine learning models to understand the data, so I imported TFIDF vectorizer and I have made the max feature as15000.



And I have split the data into two parts X and y and made them ready for training. I have created a new feature name label and summed all the other numerical column and changed the output as binary i.e., if thesentenceiscategorisedasmalignantitwillbe1orelseitwillbe0.

## Data Inputs- Logic- Output Relationships

I have analysed the input output logic with word cloud and I have wordcloudedthesentencedthatasclassifiedasfoullanguageinevery category.



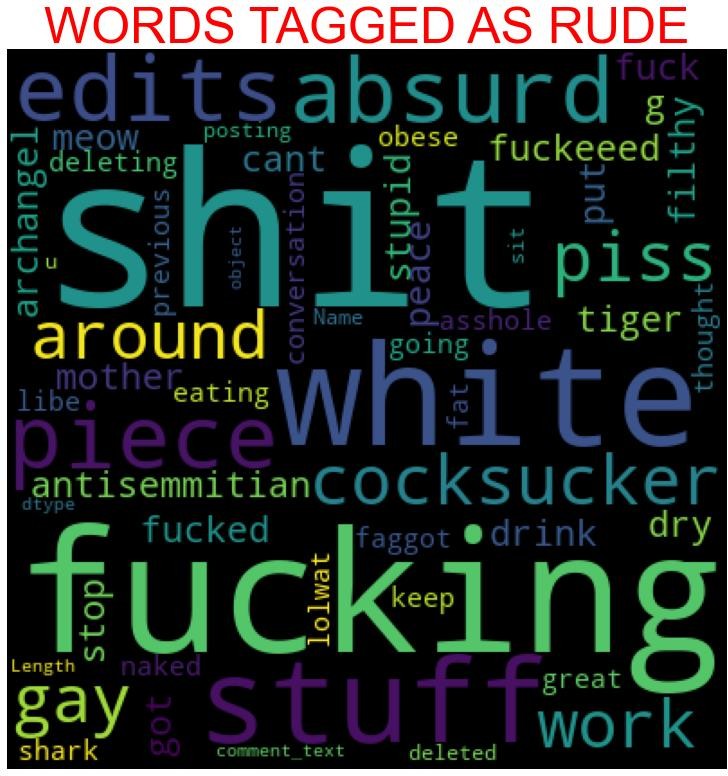
### Key observation.

We can see the foul words that are mostly used in malignant classifiedsentencesweareseeingtop400wordsthewordswhichare bigger in size are mostlyused.



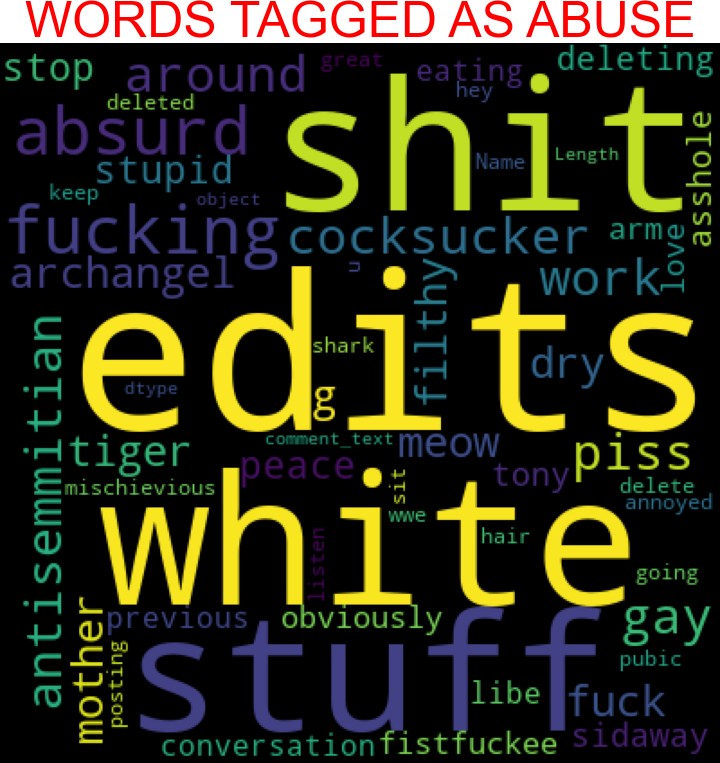
### Key observation.

Wecanseethefoulwordsthataremostlyusedinhighly\_malignant classifiedsentencesweareseeingtop400wordsthewordswhichare bigger in size are mostlyused.



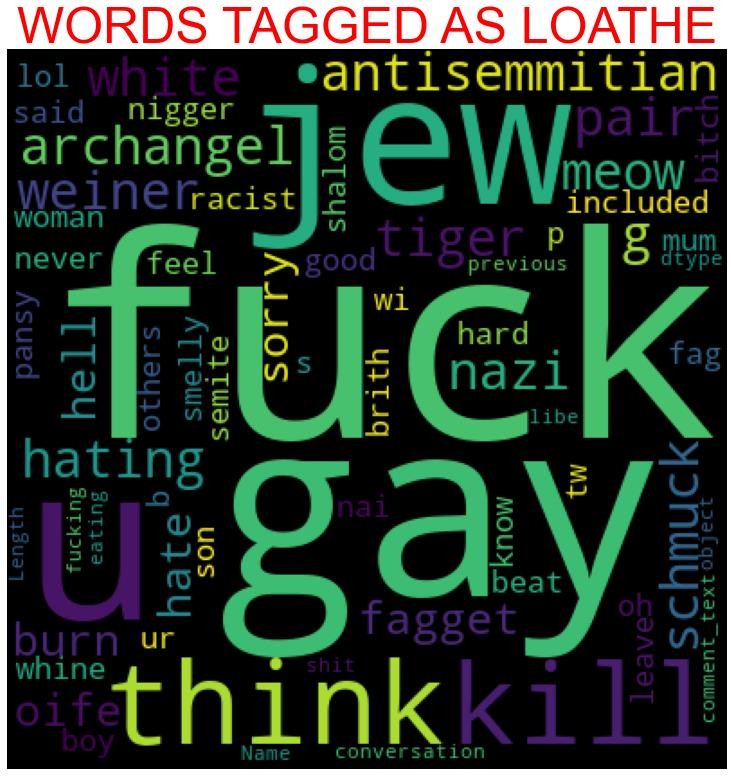
### Key observation**.**

We can see the foul words that are mostly used in rude classified sentencesweareseeingtop400wordsthewordswhicharebiggerin size are mostlyused.



### Key observation.

We can see the foul words that are mostly used in abuse classified sentences we are seeing top 400 words the words which are biggerin size are mostlyused.



### Key observation.

We can see the foul words that are mostly used in loathe classified sentencesweareseeingtop400wordsthewordswhicharebiggerin size are mostlyused.

## HardwareandSoftwareRequirementsandToolsUsed

1. Python3.8.
2. NumPy.
3. Pandas.
4. Matplotlib.
5. Seaborn. 6. Data science.
6. SciPy
7. Sklearn.
8. Anaconda Environment, JupyterNotebook.

# Model/sDevelopmentandEvaluation

## Testing of Identified Approaches (Algorithms)

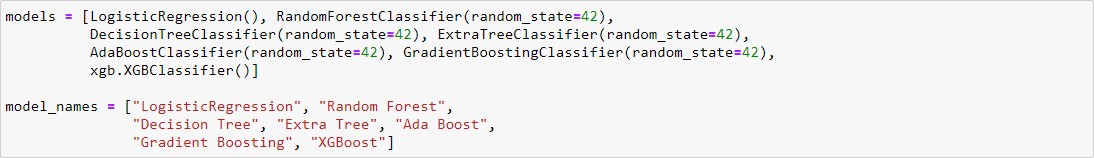
I have started the training in selecting the best random state parameter for the model as follows.

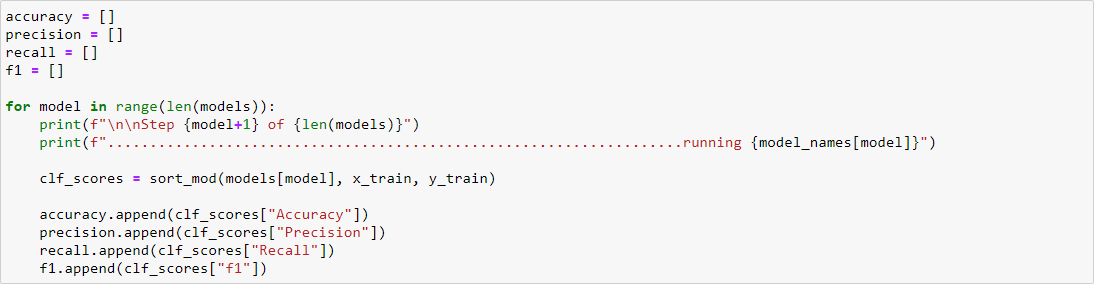


After selecting the best random state parameter, I have spitted the data into test and train with test size as 25 %. Again, I have imported the required libraries to import my ML algorithms.



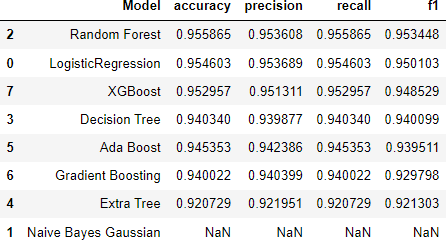
## Run and evaluate selected models

As we can see above, I have imported 8 classification algorithms andI am going to shortlist the best amongst these in basis of accuracy precision recall and F1scores.



So, like above I have run all the algorithms with the data.

## Key-Metrics for success in solving problem underconsideration.

IhavetakenkeymetricsasAccuracy,Precision,RecallandF1scores to analysis the bestmodel.

As we can see above Random Forest tops the chart, I have selected RandomForestmodelasmyfinalmodelandIhavesavedthesameforthe further usage. Further I have imported the test data and have done prediction on thesame.



# CONCLUSION

## Key Findings and Conclusions of the Study

The finding of the study is that only few users over online use unparliamentary language. And most of these sentences have more stop words, and are being long. As discussed before few motivated disrespectful crowds uses these foul languages in the online forum to bully the people around and to stop them from doing the things that they are suppose to do. Our Study helps the online forms and social media to induce a ban to profanity or usage of profanity over these forms.

## Learning Outcomes of the Study in respect of DataScience

The use of social media is the most common trend among the activities of today’s people. Social networking sites offer today’s teenagersaplatformforcommunicationandentertainment.Theyuse social media to collect more information from their friends and followers. The vastness of social media sites ensures that not all of them provide a decent environment for children. In such cases, the impact of the negative influences of social media on teenage users increases with an increase in the use of **offensive language** in social conversations.Thisincreasecouldleadto**frustration**,**depression**and a large change in their behaviour. Hence, I propose a novel approach toclassifybadlanguageusageintextconversations.Ihaveconsidered the English medium for textual conversation. I have developed our systembasedonafoullanguageclassificationapproach;itisbasedon an improved version of a Random Forest Classification Algorithm that detects offensive language usage in a conversation. As per our evaluation, we found that lesser number of users conversation is not decentallthetime.Wetrained159571observationsforeightcontext

categories using a Random Forest algorithm for context detection. Then, the system classifies the use of foul language in one of the trainedcontextsinthetextconversation.Inourtestbed,weobserved 10%ofparticipantsusedfoullanguageduringtheirtextconversation. Hence, our proposed approach can identify the impact of foul language in text conversations using a classification technique and emotion detection to identify the foul languageusage

## Limitations of this work and Scope for Future Work

The limitation of the study is that we have a imbalanced data soour model learnt more about the non-abusive sentence more than the abusive sentence. Which makes our model act like a overfit model whentestedwithlivedata.Andalso,modeltendtonotidentifyafoul or a sarcastically foullanguage.