MALIGNANT COMMENTS CLASSIFIER PROJECT PRESENTATION

SUBMITTED TO: KHUSBOO GARG
(SME OF INTERNSHIP BATCH NO-28)

SUBMITTED BY: ABHISHEK BEHERA
INTERNSHIP BATCH NO-28

INTRODUCTION

- Over a decade, social media have been growing, and people are able to express their opinions and also discuss among others via these platforms.
- These debates may arise due to differences in opinion and may often result in fights over the social media during which offensive language termed as malignant comments may be used from one side.
- This clearly pose the threat of abuse and harassment online.
- As such, some people stop giving their opinions or give up seeking different opinions which result in unhealthy and biased discussion.
- Therefore it results in different platforms and communities finding it very difficult to facilitate fair conversation and are often forced to either limit user comments or get dissolved by shutting down user comments completely.

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 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 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.

DATASET 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.

CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

- Online platforms and social media become the place where people share the thoughts freely without any partiality and overcoming all the race people share their thoughts and ideas among the crowd.
- Social media is a computer-based technology that facilitates the sharing of ideas, thoughts, and information through the building of virtual networks and communities. By design, social media is Internet-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 or applications.
- 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.
- In this huge online platform or an online community there are some people or some motivated mob wilfully bully others to make them not 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 is achieved.
- To solve this problem, we are now building a model 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.

MULTILABEL VS MULTICLASS CLASSIFICATION

As the task was to figure out whether the data belongs to zero, one or more than one categories out of the six listed in our dataset, the first step before working on the problem was to distinguish between multi-label and multi-class classification.

In multi-class classification, we have one basic assumption that our data can belong to only one label out of all the labels we have. For example, a given picture of a fruit may be an apple, orange or guava only and not a combination of these.

In multi-label classification, data can belong to more than one label simultaneously. For example, in our case a comment may be toxic, obscene and insulting at the same time. It may also happen that the comment is non-toxic and hence does not belong to any of the six labels.

Hence, I had a multi-label classification problem to solve. The next step was to gain some useful insights from data which would aid further problem solving.

DATA SCIENCE LIFE CYCLE

Data Cleaning

- Import the collected data from web scraping
- Clean and format the records as per usage by using various imputation techniques

Exploratory Data Analysis

- Check through all the dataset information like datatype, missing value, duplicate value etc.
- Analyze each and every data record to ensure we have usable information

Visualization and Data Preprocessing

- Use various visualization methods to check the data distribution identify presence of outliers and skewness
- Perform encoding and scaling methods

DATA SCIENCE LIFE CYCLE

Model Building

- Create appropriate Regression Machine Learning model function
- Need to ensure that whenever the regression function is called it is able to process all the necessary parameters

Model Evaluation

- Usage of evaluation metrics to check the accuracy of the models over trained and test data inputs
- Ensure the cross validation techniques helps in reducing over fitting and under fitting data

Hyperparameter Tuning Best Model

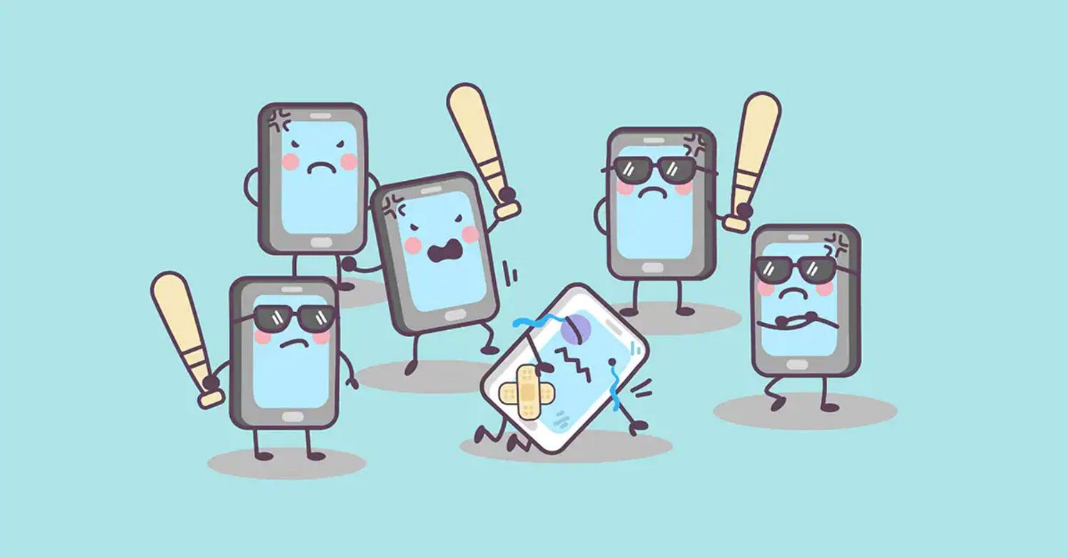
- Choosing the appropriate Regression Machine Learning model to check various parameter permutation and combinations
- Using Grid Search CV to obtain the best parameters that can be plugged into the selected model

MODEL BUILDING STEPS

1. Data Cleaning
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building
5. Model Evaluation
6. Selecting the best model

DATA PREPROCESSING

1. Load dataset
2. Remove null values
3. Drop column id
4. Convert comment text to lower case and replace '\n' with single space.
5. Keep only text data ie. a-z' and remove other data from comment text.
6. Remove stop words and punctuations
7. Apply Stemming using SnowballStemmer
8. Convert text to vectors using TfidfVectorizer
9. Load saved or serialized model
10. Predict values for multi class label



TECHNOLOGY USED

- Hardware technology being used.

RAM : 8 GB

CPU : AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz

GPU : AMD Radeon ™ Vega 8 Graphics and NVIDIA GeForce GTX 1650 Ti

- Software technology being used.

Programming language : Python

Distribution : Anaconda Navigator

Browser based language shell : Jupyter Notebook

- Libraries/Packages specifically being used.

Pandas, NumPy, matplotlib, seaborn, scikit-learn, pandas-profiling, missingno, NLTK

IMPORTED DEPENDENCIES

```
import warnings
warnings.simplefilter("ignore")
warnings.filterwarnings("ignore")
import joblib

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

import missingno
import pandas_profiling
from scipy import interp
import scikitplot as skplt
from itertools import cycle
import matplotlib.ticker as plticker

import nltk
nltk.download('stopwords', quiet=True)
nltk.download('punkt', quiet=True)
from wordcloud import WordCloud
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize, regexp_tokenize

from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV, RandomizedSearchCV
from scipy.sparse import csr_matrix

import timeit, sys
from sklearn import metrics
import tqdm.notebook as tqdm
from skmultilearn.problem_transform import BinaryRelevance
from sklearn.svm import SVC, LinearSVC
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.ensemble import AdaBoostClassifier, BaggingClassifier, RandomForestClassifier
from sklearn.metrics import hamming_loss, log_loss, accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import roc_curve, auc, roc_auc_score, multilabel_confusion_matrix
from scikitplot.metrics import plot_roc_curve
```

EXPLORATORY DATA ANALYSIS (EDA) AND VISUALIZATION

01. Univariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable.

02. Multivariate Analysis

Multivariate analysis is a set of statistical techniques used for **analysis** of data that contain more than one variable.

03. Correlation of Dataset

Correlation is used to test relationships between quantitative variables or categorical variables.

04. Correlation with Target variable

Correlation with the target variable to know how the data is related.

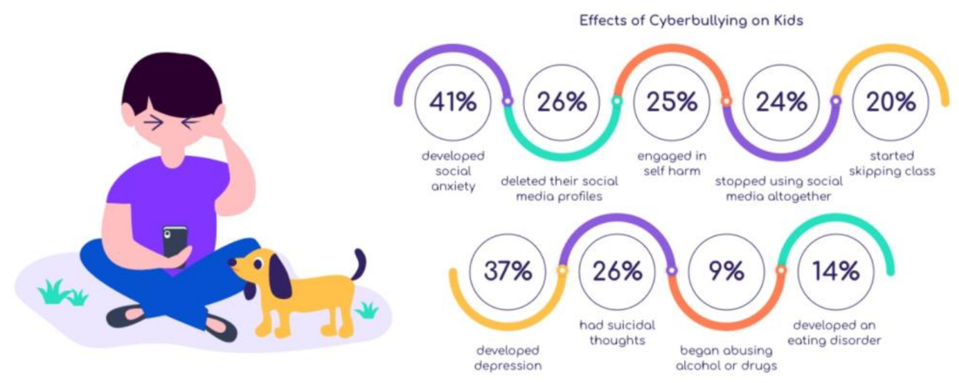
05. Conclusion

Summary with the conclusion of all the analysis

CYBERBULLYING STATISTICS

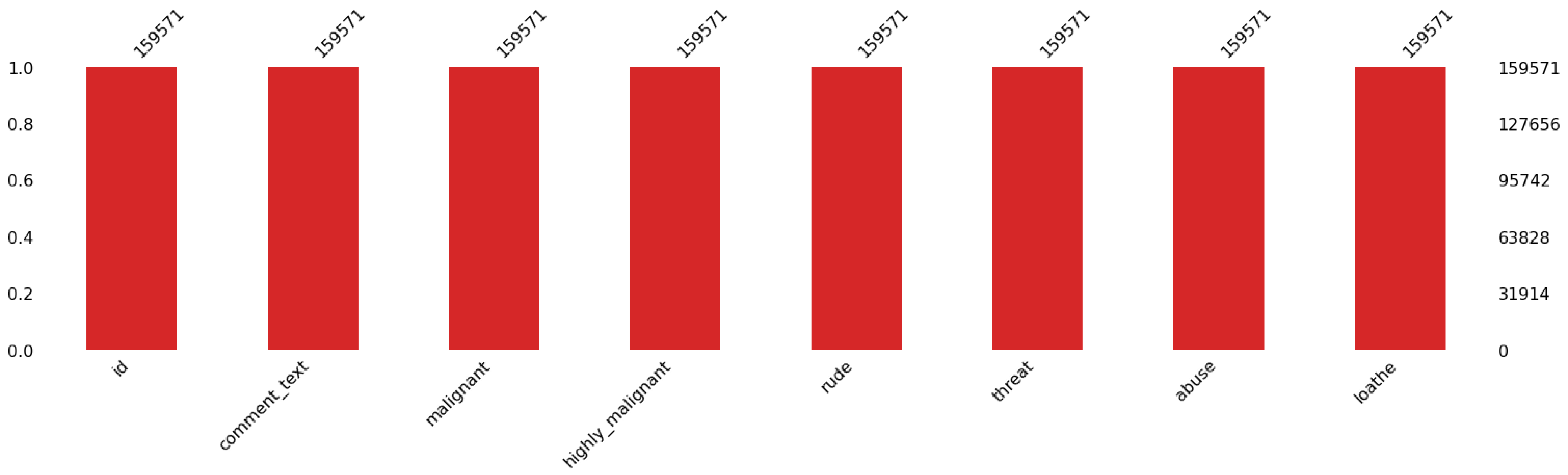
Cyberbullying has become a growing problem in countries around the world. Essentially, cyberbullying doesn't differ much from the type of bullying that many children have unfortunately grown accustomed to in school. The only difference is that it takes place online.

EFFECTS OF CYBERBULLYING

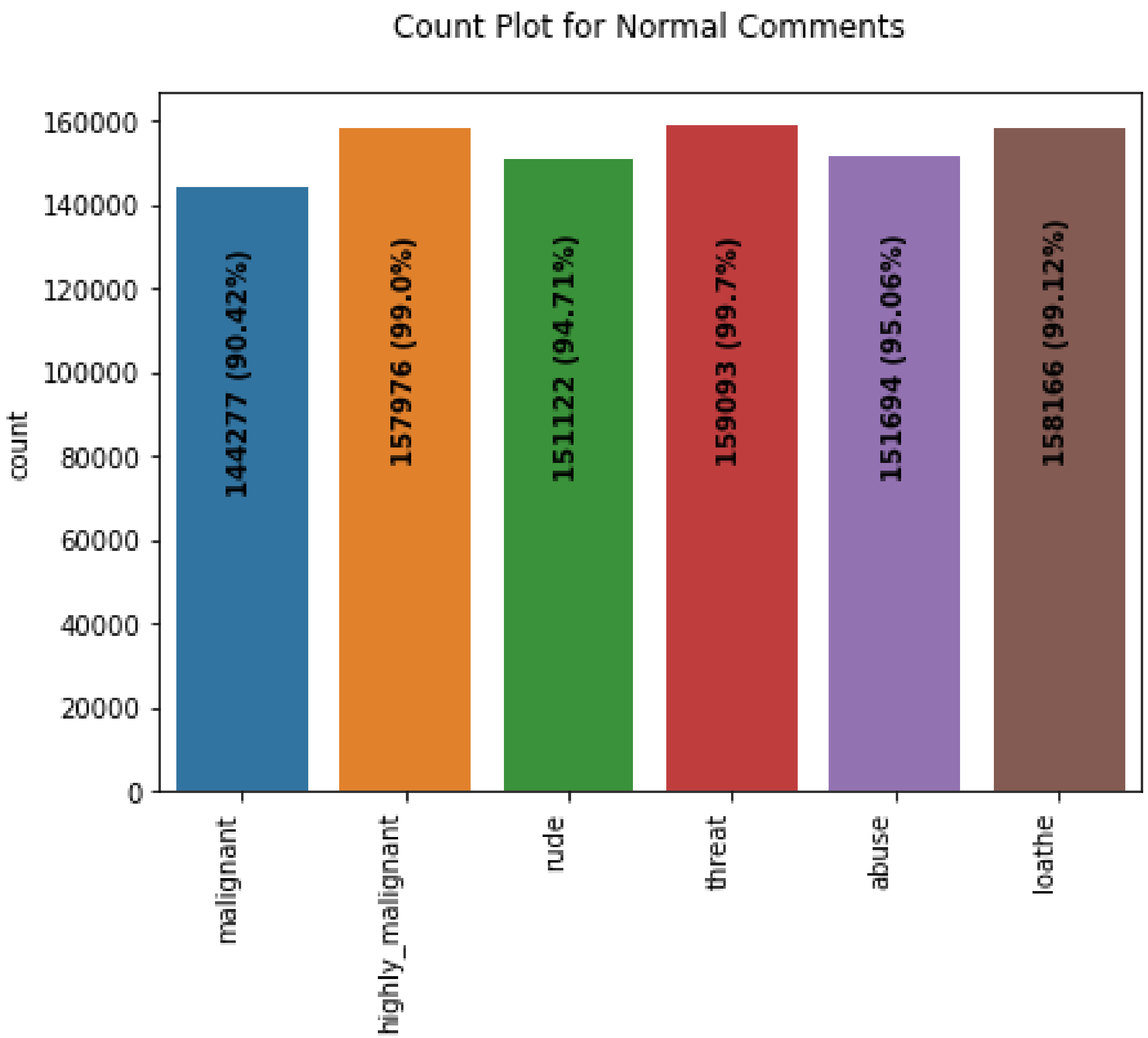


Cyberbullying is a very serious issue affecting not just the young victims, but also the victims' families, the bully, and those who witness instances of cyberbullying. However, the effect of cyberbullying can be most detrimental to the victim, of course, as they may experience a number of emotional issues that affect their social and academic performance as well as their overall mental health.

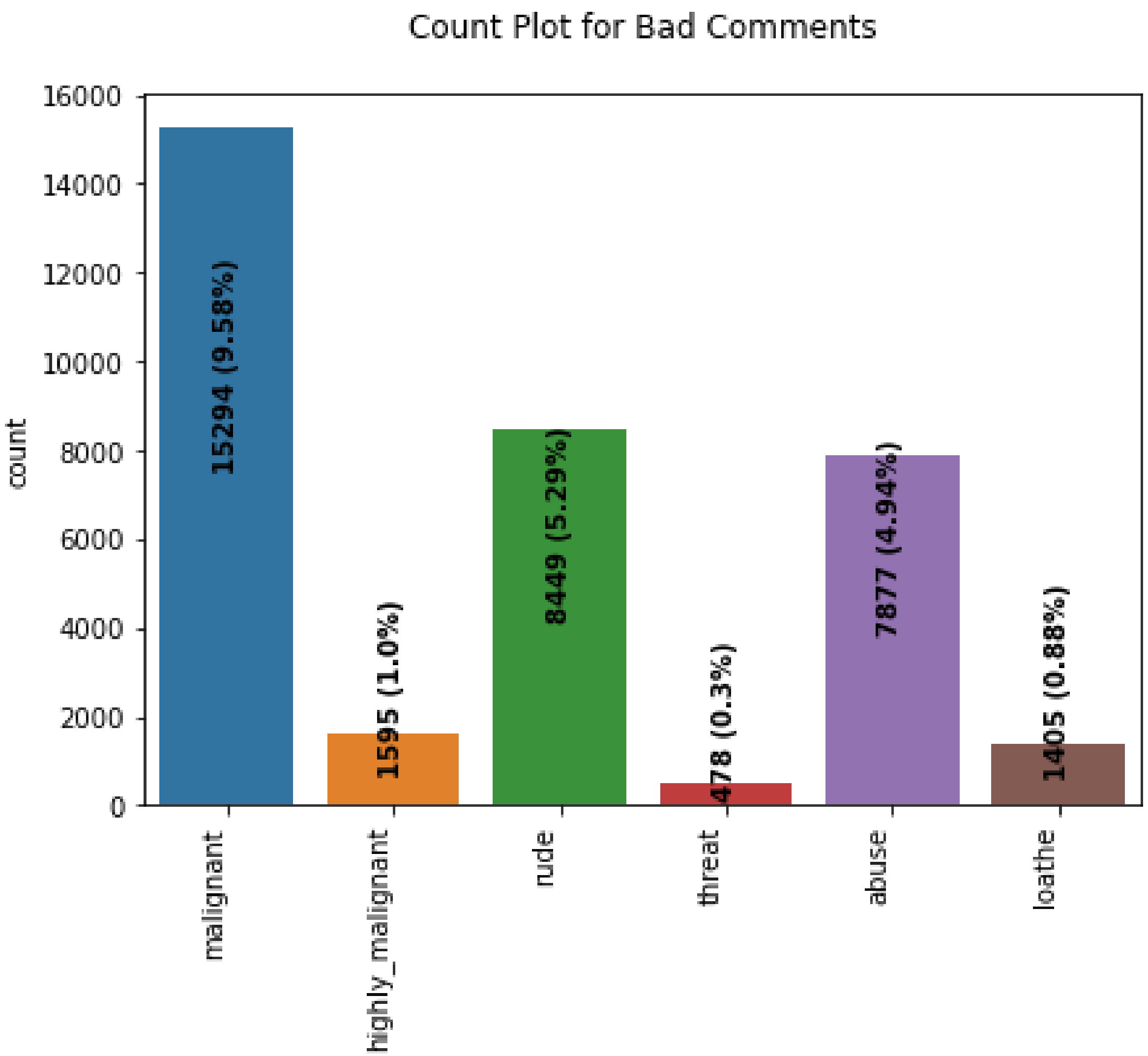
MISSING VALUES



COUNT PLOT

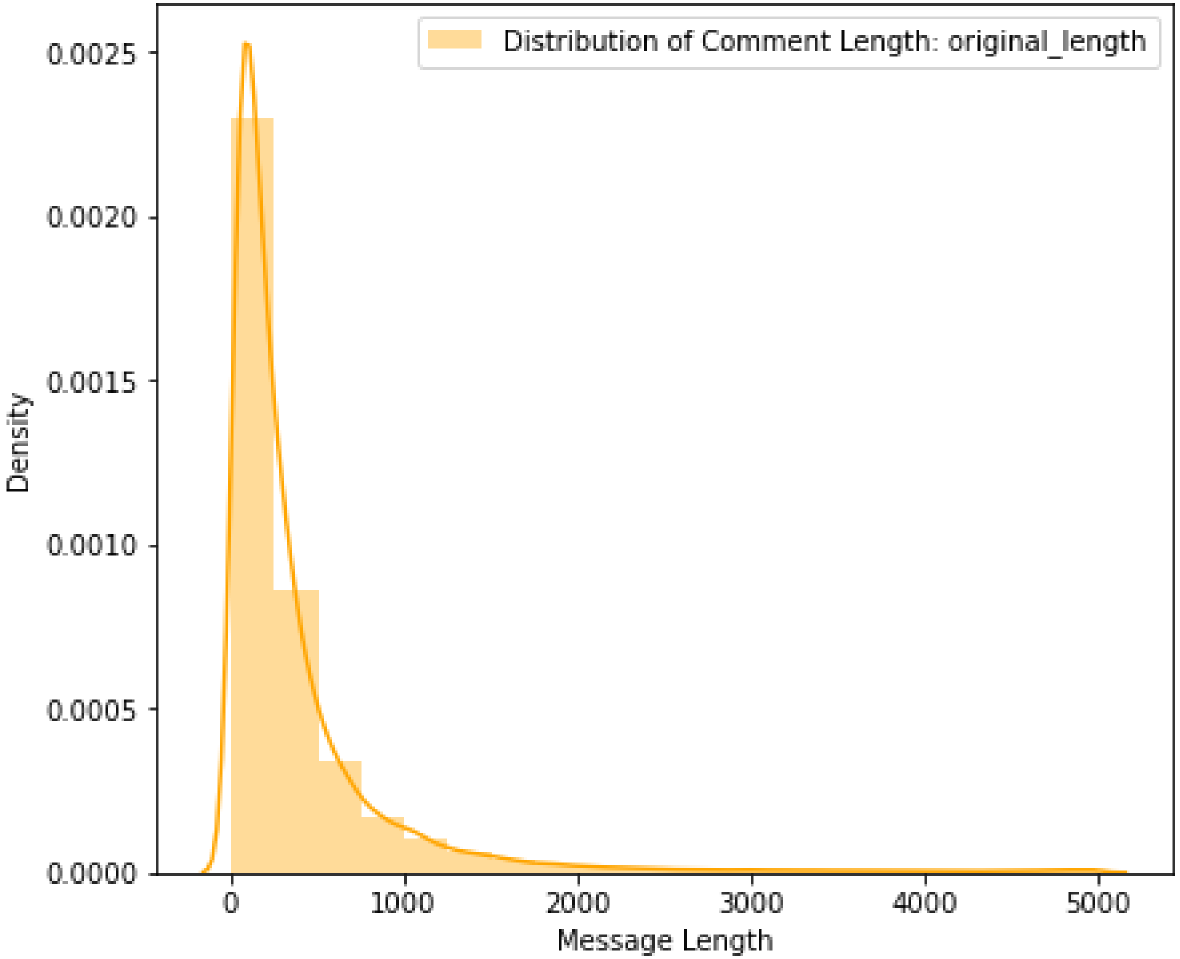


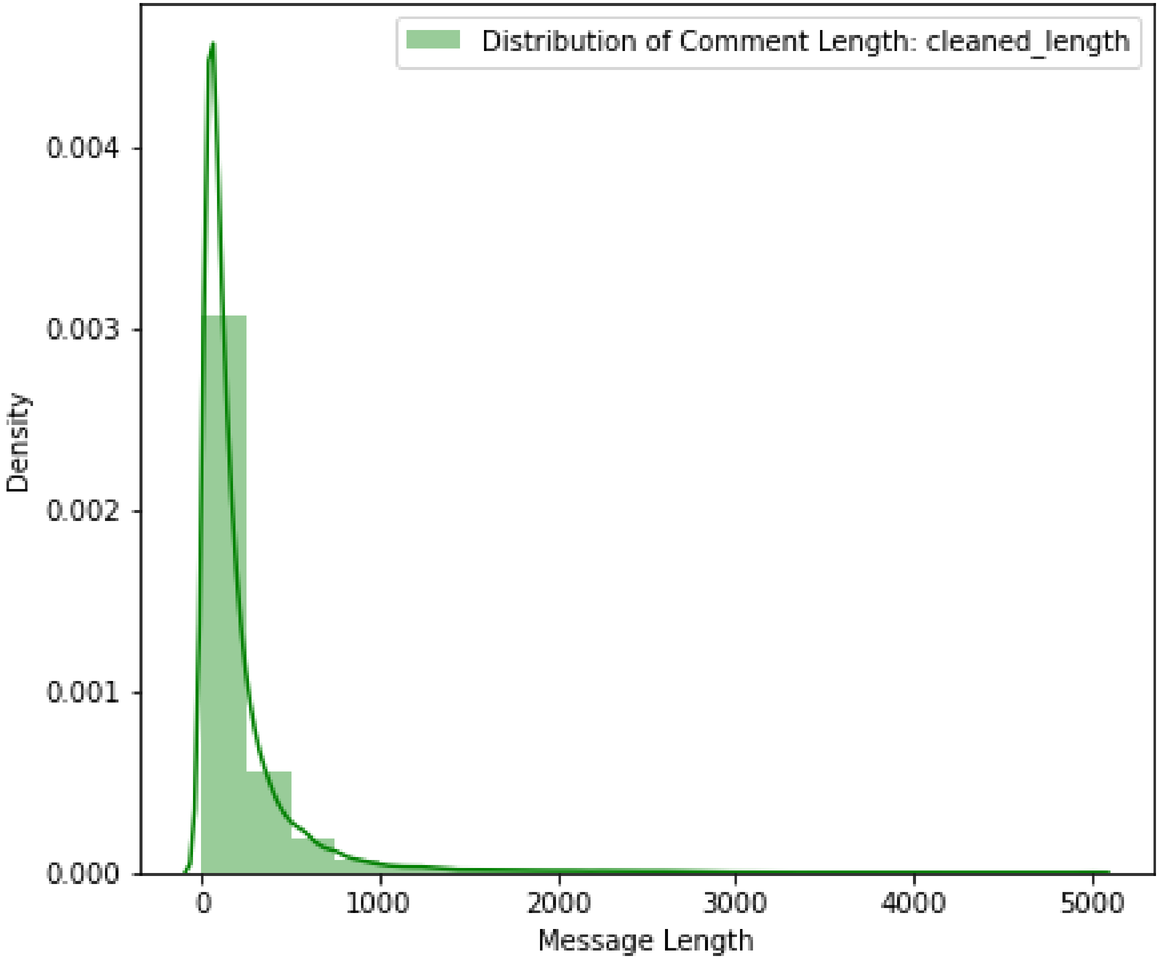
Count Plot for Normal Comments
count
160000
140000
120000
100000
80000
60000
40000
20000
0
144277 (90.42%)
157976 (99.0%)
151122 (94.71%)
159093 (99.7%)
151694 (95.06%)
158166 (99.12%)
malignant
highly_malignant
rude
threat
abuse
loathe



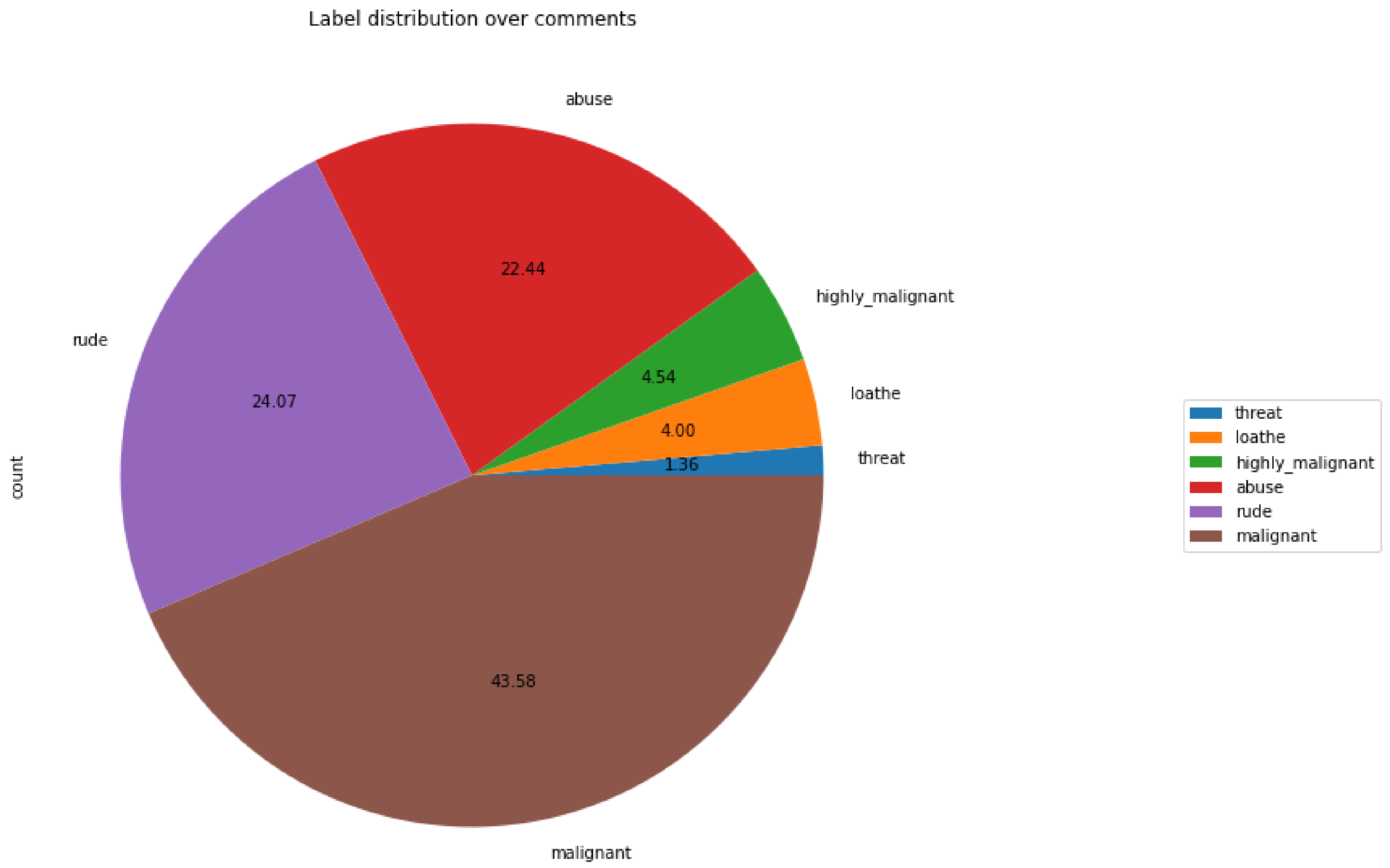
Count Plot for Bad Comments
count
16000
14000
12000
10000
8000
6000
4000
2000
0
15294 (9.58%)
1595 (1.0%)
8449 (5.29%)
478 (0.3%)
7877 (4.94%)
1405 (0.88%)
malignant
highly_malignant
rude
threat
abuse
loathe

DISTRIBUTION PLOT





PIE PLOT

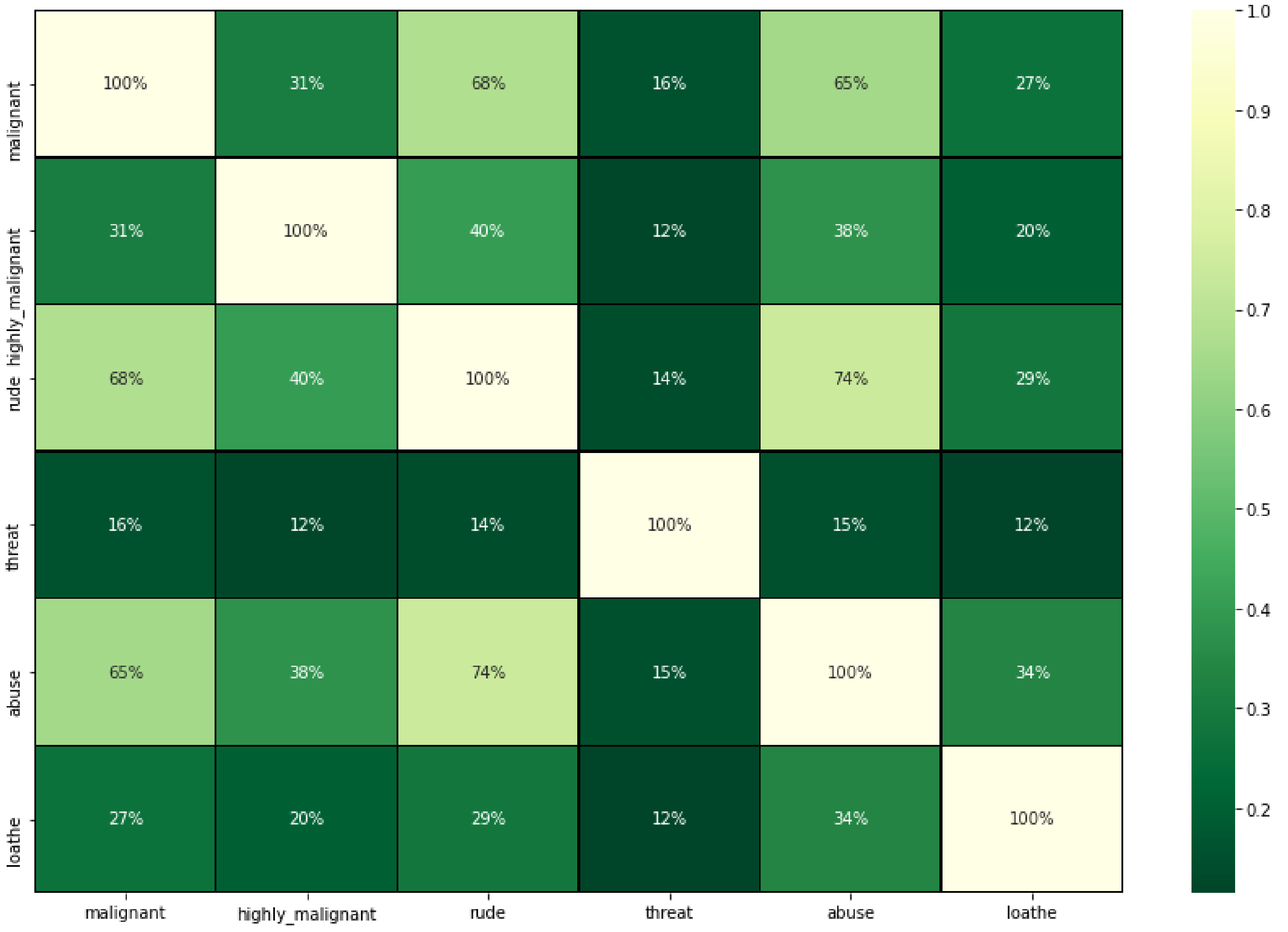


Label distribution over comments
abuse
22.44
highly_malignant
4.54
loathe
4.00
threat
1.36
rude
24.07
count
43.58
malignant
threat
loathe
highly_malignant
abuse
rude
malignant

WORD CLOUD

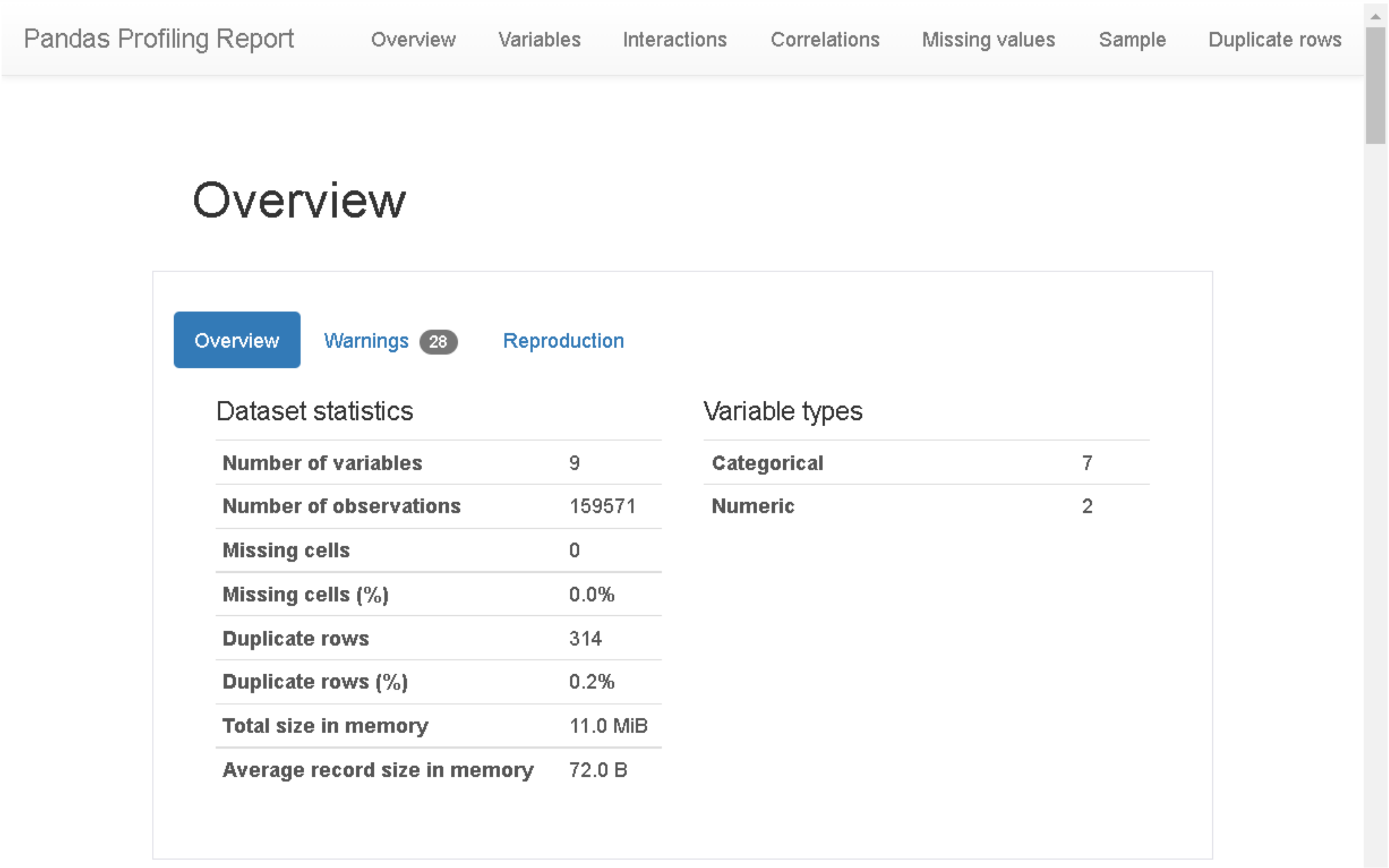
WordCloud: Representation of Loud words in BAD COMMENTS

HEATMAP



malignant
highly_malignant
rude
threat
abuse
loathe
100%
31%
68%
16%
65%
27%
31%
100%
40%
12%
38%
20%
68%
40%
100%
14%
74%
29%
16%
12%
14%
100%
15%
12%
65%
38%
74%
15%
100%
34%
27%
20%
29%
12%
34%
100%
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

PANDAS PROFILING



Pandas Profiling Report   Overview   Variables   Interactions   Correlations   Missing values   Sample   Duplicate rows

Overview

Overview   Warnings 28   Reproduction

Dataset statistics

Number of variables	9
Number of observations	159571
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	314
Duplicate rows (%)	0.2%
Total size in memory	11.0 MiB
Average record size in memory	72.0 B

Variable types

Categorical	7
Numeric	2

CLASSIFICATION FUNCTION

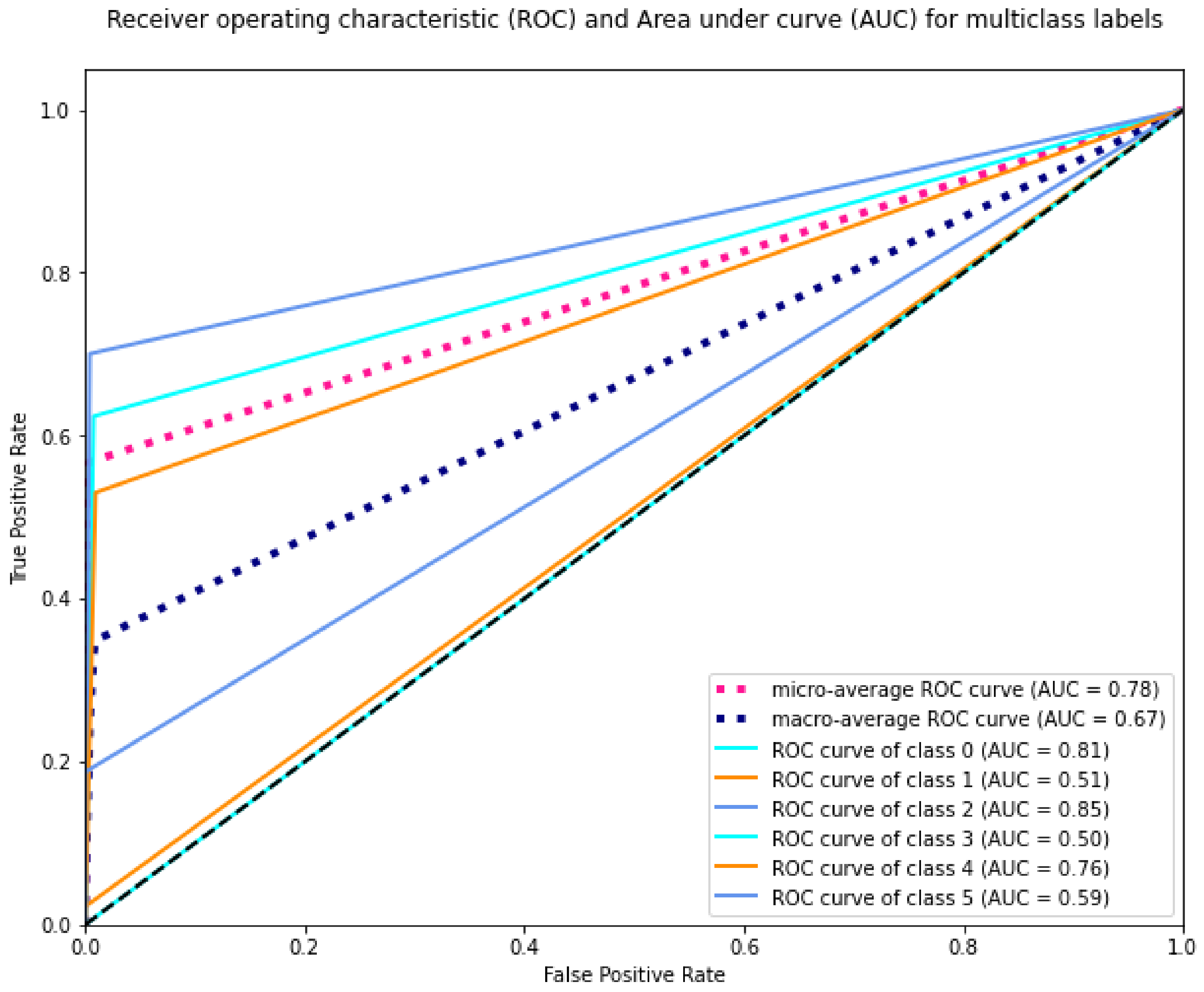
```python
# 3. Training and Testing Model on our train dataset

# Creating a function to train and test model
def build_models(models,x,y,test_size=0.33,random_state=42):
    # spliting train test data using train_test_split
    x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=test_size,random_state=random_state)

    # training models using BinaryRelevance of problem transform
    for i in tqdm.tqdm(models,desc="Building Models"):
        start_time = timeit.default_timer()

        sys.stdout.write("\n==========================================================================================\n")
        sys.stdout.write(f"Current Model in Progress: {i} ")
        sys.stdout.write("\n==========================================================================================\n")

        br_clf = BinaryRelevance(classifier=models[i]["name"],require_dense=[True,True])
        print("Training: ",br_clf)
        br_clf.fit(x_train,y_train)

        print("Testing: ")
        predict_y = br_clf.predict(x_test)

        ham_loss = hamming_loss(y_test,predict_y)
        sys.stdout.write(f"\n\tHamming Loss  : {ham_loss}")

        ac_score = accuracy_score(y_test,predict_y)
        sys.stdout.write(f"\n\tAccuracy Score: {ac_score}")

        cl_report = classification_report(y_test,predict_y)
        sys.stdout.write(f"\n{cl_report}")

        end_time = timeit.default_timer()
        sys.stdout.write(f"Completed in [{end_time-start_time} sec.]")

        models[i]["trained"] = br_clf
        models[i]["hamming_loss"] = ham_loss
        models[i]["accuracy_score"] = ac_score
        models[i]["classification_report"] = cl_report
        models[i]["predict_y"] = predict_y
        models[i]["time_taken"] = end_time - start_time

        sys.stdout.write("\n==========================================================================================\n")

    models["x_train"] = x_train
    models["y_train"] = y_train
    models["x_test"] = x_test
    models["y_test"] = y_test

    return models
```

```python
# Preparing the list of models for classification purpose
models = {"GaussianNB": {"name": GaussianNB()},
          "MultinomialNB": {"name": MultinomialNB()},
          "Logistic Regression": {"name": LogisticRegression()},
          "Random Forest Classifier": {"name": RandomForestClassifier()},
          "Support Vector Classifier": {"name": LinearSVC(max_iter = 3000)},
          "Ada Boost Classifier": {"name": AdaBoostClassifier()},
          "K Nearest Neighbors Classifier": {"name": KNeighborsClassifier()},
          "Decision Tree Classifier": {"name": DecisionTreeClassifier()},
          "Bagging Classifier": {"name": BaggingClassifier(base_estimator=LinearSVC())},
         }

# Taking one forth of the total data for training and testing purpose
half = len(df)//4
trained_models = build_models(models,X[:half,:],Y[:half,:])
```

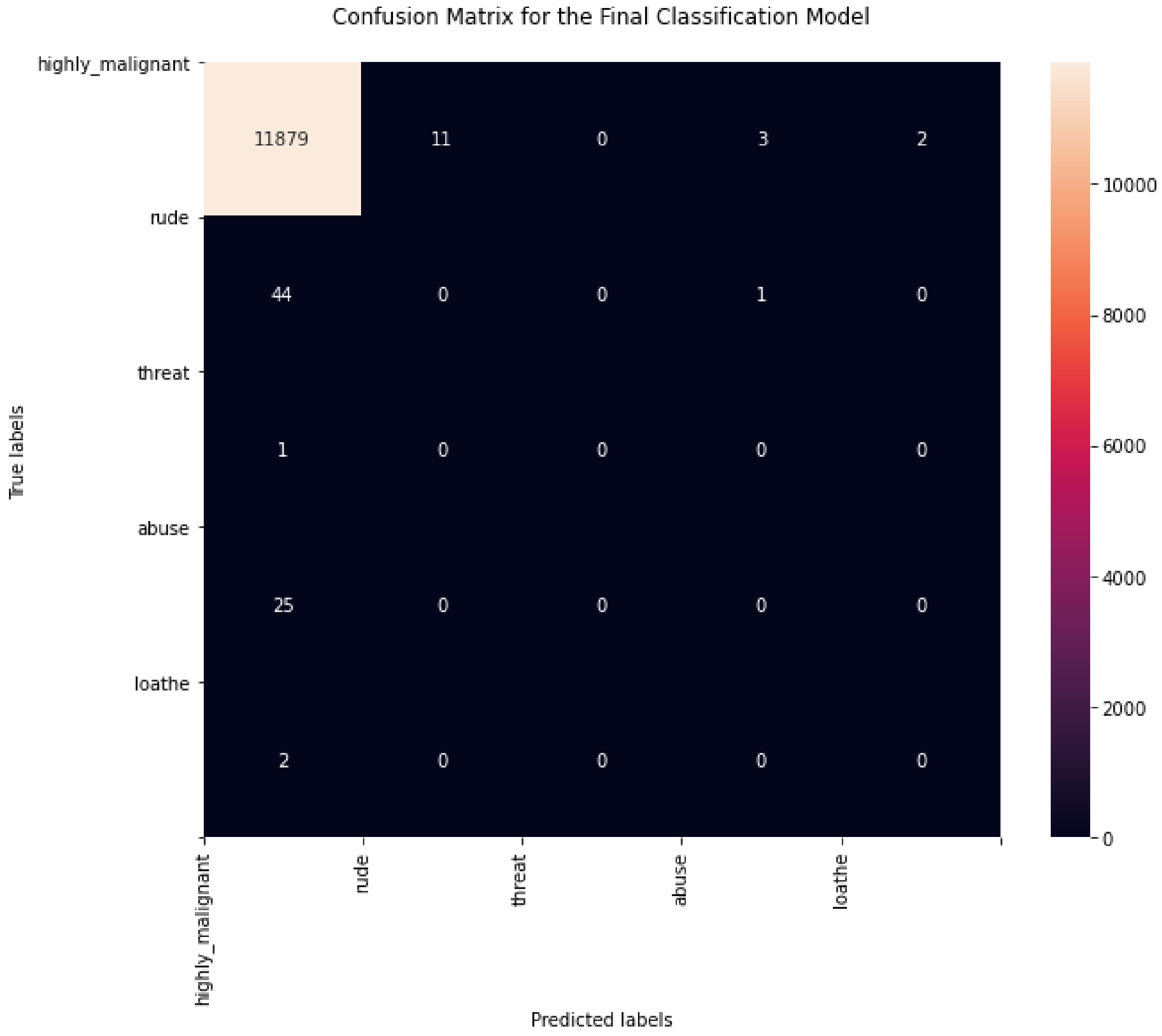
CLASSIFICATION MACHINE LEARNING MODELS

```
Building Models: 100%  9/9 [1:26:58<00:00, 756.96s/it]

=====================================================================================
Current Model in Progress: GaussianNB
=====================================================================================
Training:  BinaryRelevance(classifier=GaussianNB(), require_dense=[True, True])
Testing:

        Hamming Loss  : 0.21560957083175086
        Accuracy Score: 0.4729965818458033
              precision    recall  f1-score   support

           0       0.16      0.79      0.26      1281
           1       0.08      0.46      0.13       150
           2       0.11      0.71      0.19       724
           3       0.02      0.25      0.03        44
           4       0.10      0.65      0.17       650
           5       0.04      0.46      0.07       109

   micro avg       0.11      0.70      0.20      2958
   macro avg       0.08      0.55      0.14      2958
weighted avg       0.12      0.70      0.21      2958
 samples avg       0.05      0.07      0.05      2958
Completed in [27.41599629999999 sec.]
=====================================================================================
```

ROC AUC CURVE



CONFUSION MATRIX



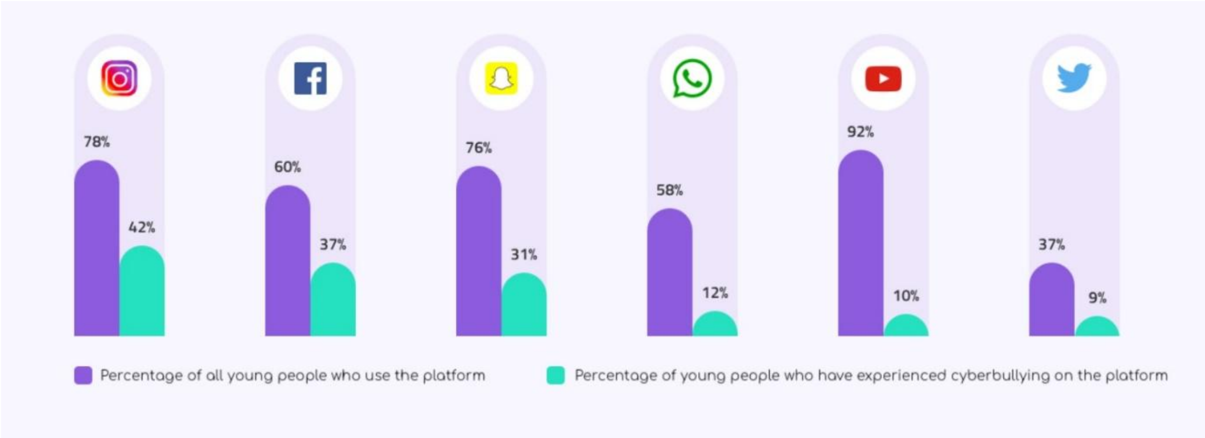
KEY FINDINGS AND CONCLUSIONS OF THE STUDY

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 quite long.

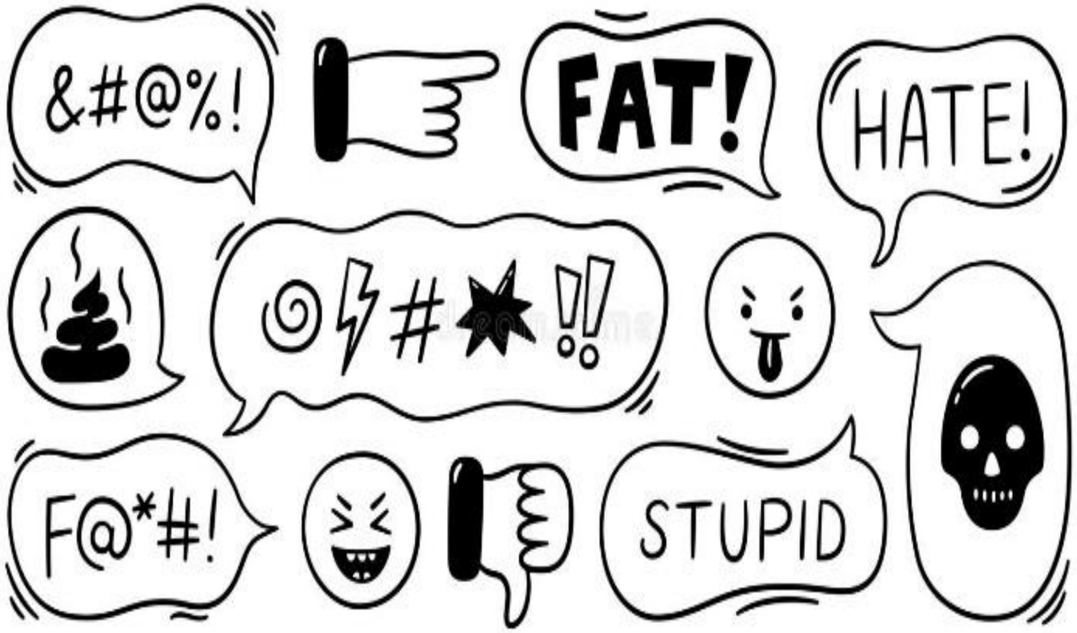
As discussed before few motivated disrespectful crowds use these foul languages in the online forum to bully the people around and to stop them from doing these things that they are not supposed to do.

Our study helps the online forums and social media to induce a ban to profanity or usage of profanity over these forums.



LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

Through this project we were able to learn various Natural language processing techniques like lemmatization, stemming, removal of stop words. We were also able to learn to convert strings into vectors through hash vectorizer. In this project we applied different evaluation metrics like log loss, hamming loss besides accuracy.



LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

My point of view from my project is that we need to use proper words which are respectful and also avoid using abusive, vulgar and worst words in social media. It can cause many problems which could affect our lives. Try to be polite, calm and composed while handling stress and negativity and one of the best solutions is to avoid it and overcoming in a positive manner.

LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

Problems faced while working in this project:

- More computational power was required as it took more than 2 hours
- Imbalanced dataset and bad comment texts
- Good parameters could not be obtained using hyperparameter tuning as time was consumed more

Areas of improvement:

- Could be provided with a good dataset which does not take more time.
- Less time complexity
- Providing a proper balanced dataset with less errors.

THANK YOU

ABUSE
BAD
BEHAVIOR
OBJECTION
AGAINST
HARASSMENT
HABIT
FORCE
PROTEST
SOCIAL
TRAUMA
BEHAVIOR
HABIT
PROBLEM
AGAINST
HOME
FORCE
PROTEST
HARASSMENT
EMOTIONAL
VICTIM
PROTECTION
WORKPLACE
ASSAULT
SCHOOL
GROUP
RELATIONAL
DOMINATION
MENTAL
DESPAIR
COERCION
CYBER
MOBBING
CYBER
BULLYING
HOME
SCARED
AGGRESSIVE
NEIGHBORHOODS
VIOLENCE
GROUP
PREVENTION
THREAT
COERCION
WORKPLACE
STRESS
GROUP