FLIP ROBO

# Malignant comments

Submitted by:

Ria Maitra

# Abstract

The Internet has allowed people across the world to connect instantaneously and has revolutionised the way we communicate and share information with one another. More than 4 billion people were Internet users in 2018, more than half of the global population.

In many ways, the Internet has had a positive influence on society. For example, it helps us to communicate easily and to share knowledge on all kinds of important topics efficiently: from the treatment of disease to disaster relief. But the Internet has also broadened the potential for harm. Hateful messages and incitements to violence are distributed and amplified on social media in ways that were not previously possible.

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.

In this paper we have examined the comments column. The text data has been pre-processed and with the help of tf-idf all text data has been converted into vectors.

Our data contains seven columns comments, malignant, highly malignant, rude, abuse, loathe, threat. After doing exploratory analysis on comments column, it is been observed malignant comments are based on racism, sexism and life threat words.

We observed build three models, MultinomialNB, Decision Three and Knn models on each of our target variables. MultinomialNB is the best performing model, hence we have used it on our test data to for prediction.

# Introduction

Worldwide accessibility to the Internet has incredibly reshaped our perception of the world. One of the children of the World Wide Web is Social Media (SM), which is present in many forms: online game platforms, dating apps, forums, online news services, and social networks. Different social networks aim at different objectives: opinion transmission (Twitter or Facebook), business contacts (LinkedIn), image sharing (Instagram), video transmission (YouTube), dating (Meetic), and so on. However, they all have one thing in common: they aim to connect people. The power of social networking is so great that the number of worldwide users is expected to reach 3.02 billion active social media users per month by 2021. This will account for approximately one-third of the Earth's population.

In recent years, social networks (and especially Twitter) have been used to spread hate messages.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness, insults personal attacks, provocation, racism, sexism, threats, or toxicity, has been identified as a major threat on online social media platforms. Hate speech refers to a kind of speech that denigrates a person or multiple persons based on their membership to a group, usually defined by race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views. The Rabat Plan of Action of the United Nations , which defines the guidelines to distinguish between free speech and hate speech, recommends differentiating between three types of expressions: "expression that constitutes a criminal offence; expression that is not criminally punishable, but may justify a civil suit or administrative sanctions; expression that does not give rise to criminal, civil or administrative sanctions, but still raises concern in terms of tolerance, civility and respect for the rights of others."

Related to this, hate crimes are a type of violation of the law whose primary motivation is the existence of prejudices regarding the victims. This occurs when the offender chooses victims on grounds that they belong to a certain group defined basically by the attributes mentioned earlier. There is evidence that hate crimes are influenced by singular widely publicized events (terrorist attacks, uncontrolled migration, demonstrations, riots, etc.). These events usually act as triggers, and their effect is dramatically increased inside SM. This makes SM a sensor in the real world and a source of valuable information for crime forecasting. In fact, social networks are filled with messages from individuals inciting punishment against different targeted groups. When these messages are collected after a trigger event over a period of time, they can be used for the analysis of hate crimes in all the phases: climbing, stabilization, duration, and decline of the threat. Therefore, monitoring SM becomes a priority for the forecasting, detection and analysis of hate crimes.

Pew Research Center reports that among 4248 adults in the United States, 41% have personally experienced harassing behavior online, whereas 66% witnessed harassment directed towards others.

Around 22% of adults have experienced offensive name-calling, purposeful embarrassment (22%), physical threats (10%), and sexual harassment (6%), among other types of harassment. Social media platforms are the most prominent grounds for such toxic behavior. Even though they often provide ways of flagging offensive and hateful content, only 17% of all adults have flagged harassing conversation, whereas only 12% of adults have reported someone for such acts. Manual techniques like flagging are neither effective nor easily scalable and have a risk of discrimination under subjective judgments by human annotators. Since an automated system can be faster than human annotation, machine learning models to automatically detect online hate have been gaining popularity and bringing researchers from different fields together.

## 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.

## Review of Literature:

Theoretical underpinnings of online hate

Several concepts are commonly associated with the definition of online hate in the literature. As a phenomenon, online hate is cross-disciplinary; it has been studied using multiple theoretical lenses and conceptual frameworks, including social psychology, Human–Computer Interaction, politics, and legislation/regulative aspects. For example, Kansara et al present a framework for cyberbullying in social networks that contains harassment (i.e., sending offensive text messages and images), flaming (online violence using harsh messages), outing (personal information dissemination), exclusion (singling or leaving someone out of group), and masquerading (offensive communication using Sybil identities). Marret and Choo present a framework of online victimization that highlights offline perpetration and parental conflict. These studies highlight the complex dynamics of online hate that complicate its automatic detection.

Figure 1 displays our conceptual framework of the focus areas in the extant OHR. First, online is seen as the use of abusive, offensive, or profane language. These studies tend to focus on the language aspects of online hate, such as linguistic styles, vocabularies, and ways of expression. Some of these studies deal with "counterspeech", i.e., ways of defusing the hateful comments with language-based strategies

Second, some studies focus particularly on online hate as hate speech i.e., "offensive post, motivated, in whole or in part, by the writer's bias against an aspect of a group of people. [underlining by us]". The focal dimension here is targeting; i.e., the hate has a specific target such as refugees, women, a race, or religion. Waseem et al. ] distinguish between different types of abuse segmented by the target of the abuse directed towards an individual/entity or generalized towards a group and the degree to which it is explicit. ElSherief et al. study the relationship of hate instigators and targets and online visibility, finding that high-profile social media users attract more hate. Salminen et al. find media and police to be major targets of hate in online news commenting. Overall, news-related discussions have been considered as a major hotbed for online toxicity

Third, another important aspect of OHR is the consideration of group dynamics, visible in the studies focused on online hate groups and group prejudice, persuasive storytelling as hate conditioning radicalization via social media extremist content, cultural transmission of hate, social exclusion, and so on. Due to a high degree of contextual and subjective factors, these nuances are often studied using interpretative methods.

Fourth, some studies focus on the consequences of online hate, meaning its effects on individuals and groups, for example, on the health of social media communities. Often, these studies involve a predictive machine learning aspect for the detection and classification of toxicity in specific communities and social media platforms. The central characteristic of toxicity studies is that they perceive online hate not only as the use of language but also as an action having a concrete effect or outcome. These outcomes may include the user leaving the toxic discussion, "silencing" or reduced participation in online social media, radicalization, group polarization where the previously held prejudices are enforced, degraded quality (“health”) of an online community, offline violence and security threats, and decreased feelings of safety and wellbeing of online users.

Finally, computer science studies in this field tend to focus on automating the detection of online hate. The positioning of this research falls within the computational stream of research, meaning experimentation with classifiers and features to improve automatic hate detection.

Online hate is composed of the use of language that contains either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity – in other words, comments that are rude, disrespectful, and can result in negative online and offline consequences for the individual, community, and society at large.

## Objectives:

- To find out the most used abusive words in the comment section.
- To build models to classify type of comments

## Significance of the Study:

This study will help to understand the type comments and to understand the most abusive words used. This will help to monitor comments in social media, hence preventing any kind of mental damage to other people emotions.

# Methodology

This research has used:

- Univariate analysis
- Natural Processing Languages and
- Machine Learning Algorithms

To study the relationship between independent variable and dependent variables

- Machine learning Algorithms: to classify fake news

**Research Aim:**

The research aim is to build a model to classify malignant, highly malignant, rude, threat, loathe, abusive comments.

**Research Design:**

The research has used the following steps:

- Exploratory Data Analysis
- Text Cleaning
- Removing spaces, punctuations, special characters
- Tf-idf: to convert texts into vectors
- Building Machine learning Models
- Conclusion

**Data Collection:**

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.

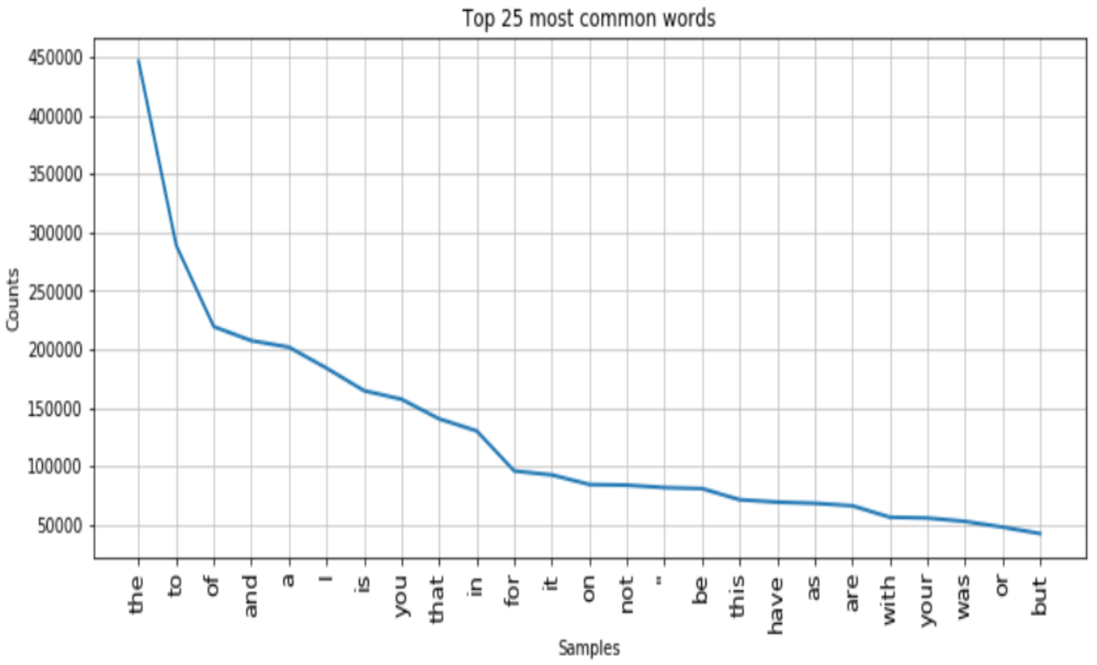
This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes many categories of comments, we can do good amount of data exploration and derive some interesting features using the comments text column available.

You need to build a model that can differentiate between comments and its categories.

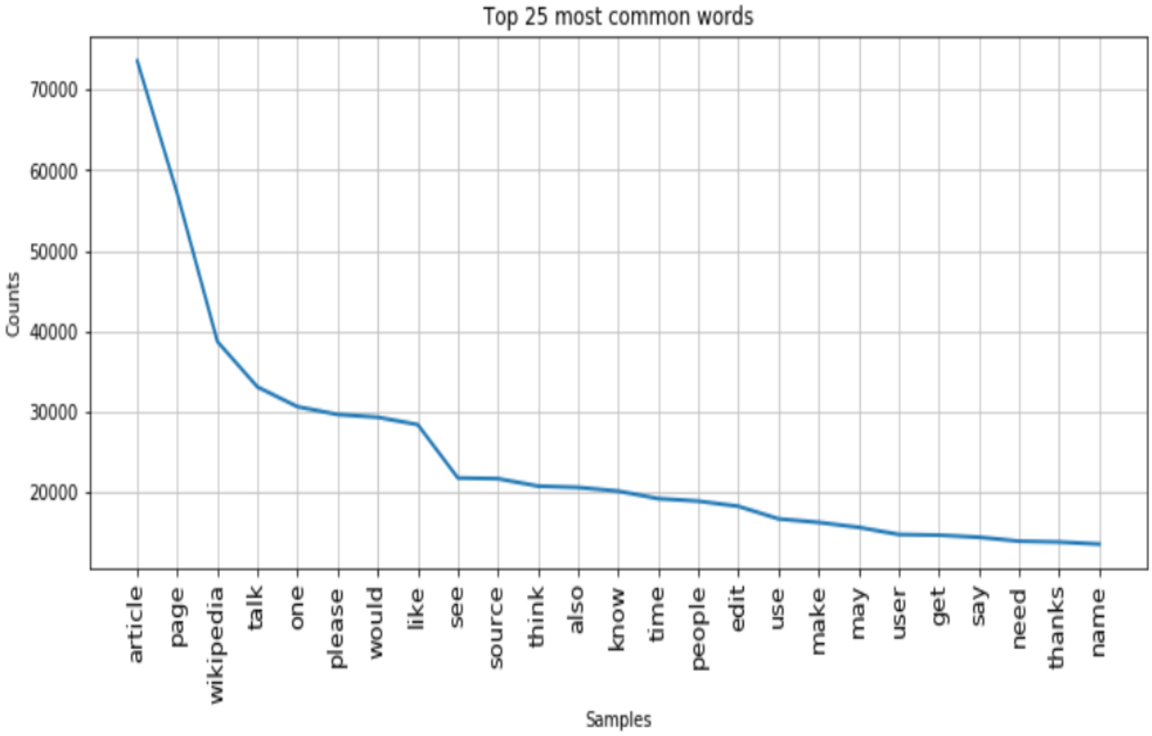
Refer to the data set file provided along with this.

## Data Preprocessing:

### Checking frequency of words before cleaning the text comment

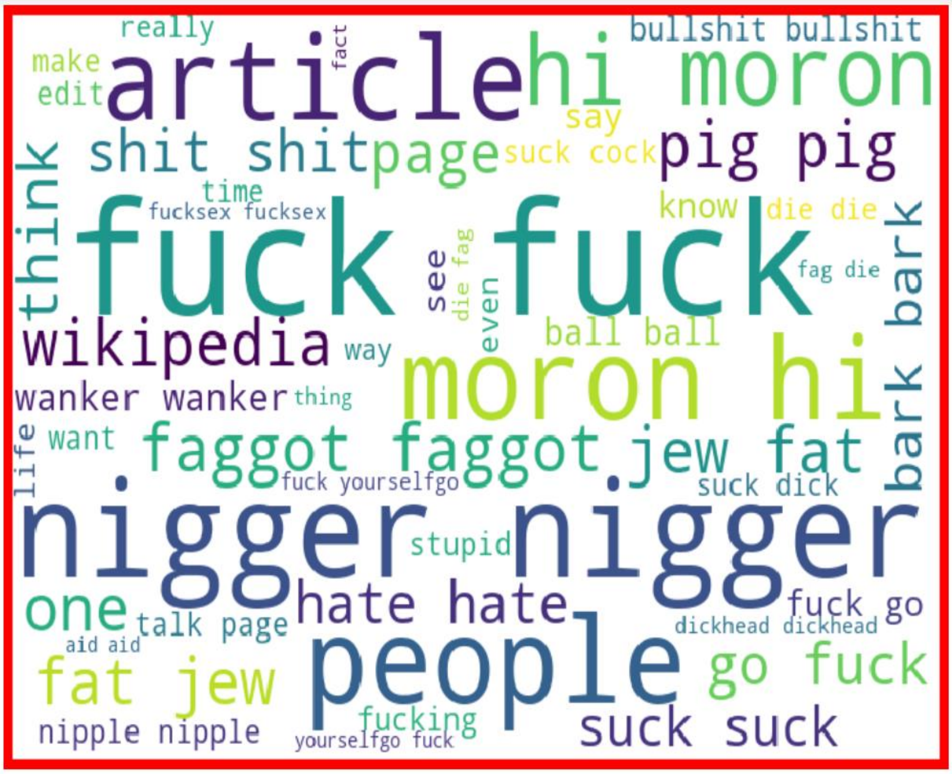


### Checking frequency of words after cleaning the text comment

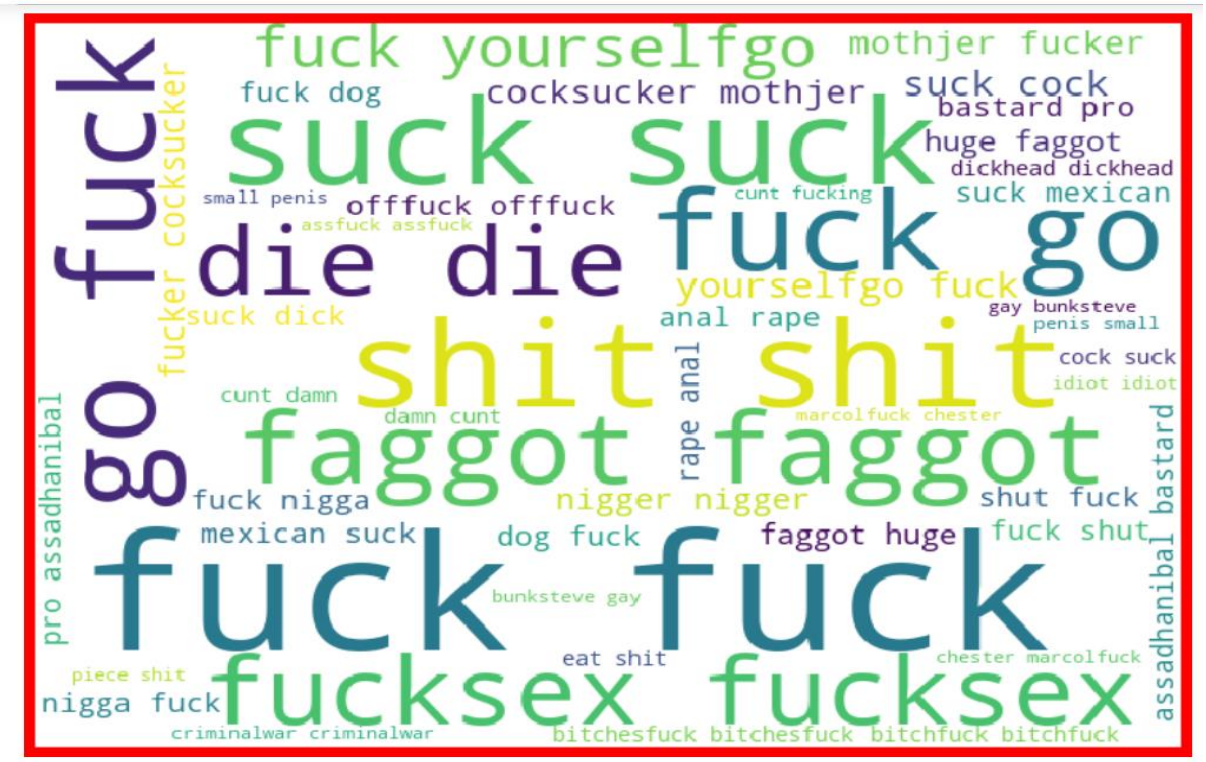


- Checking loud words in comments column

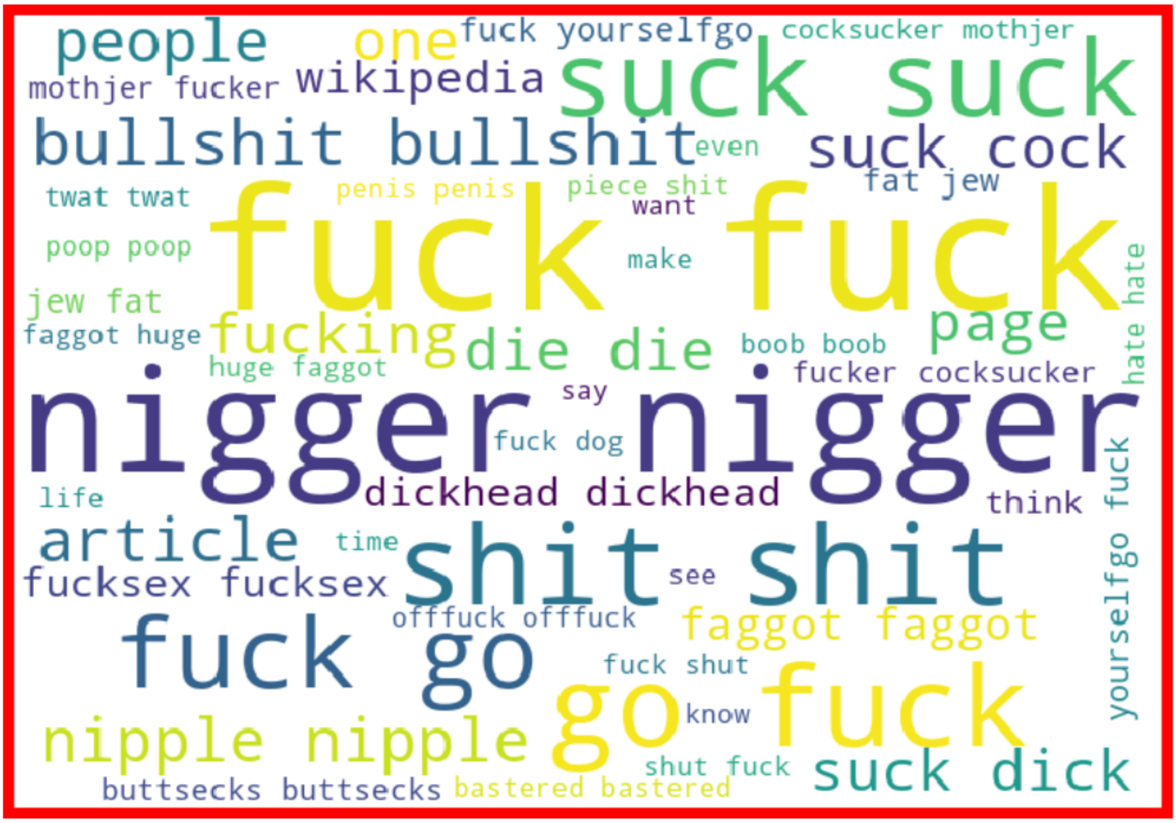
**Malignant**



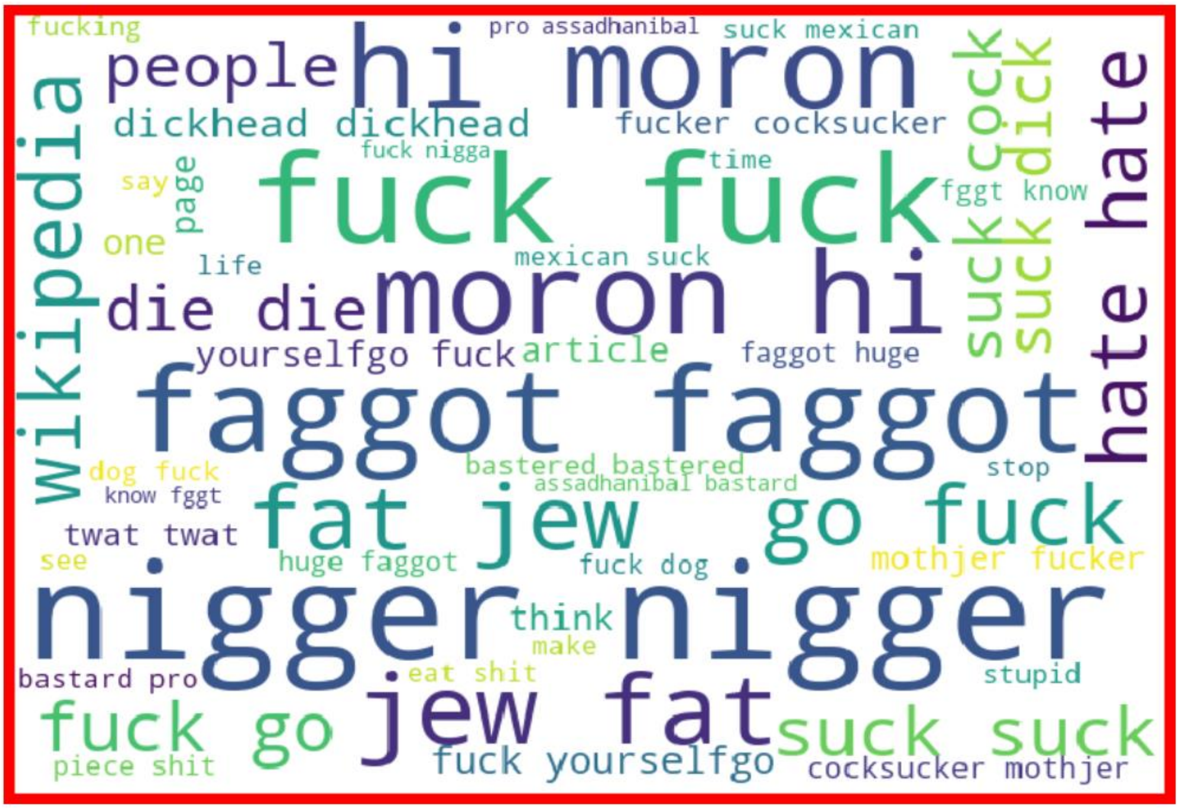
**Highly Malignant**



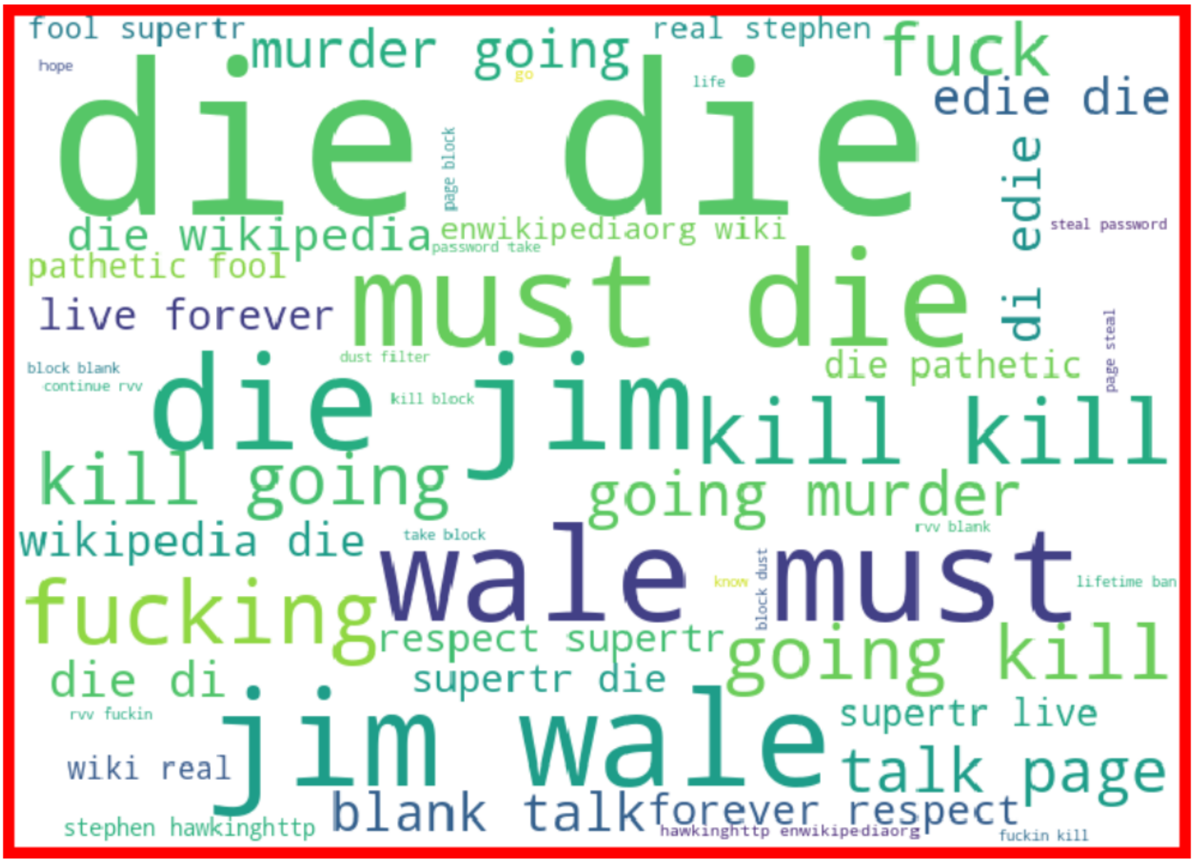
**Rude**



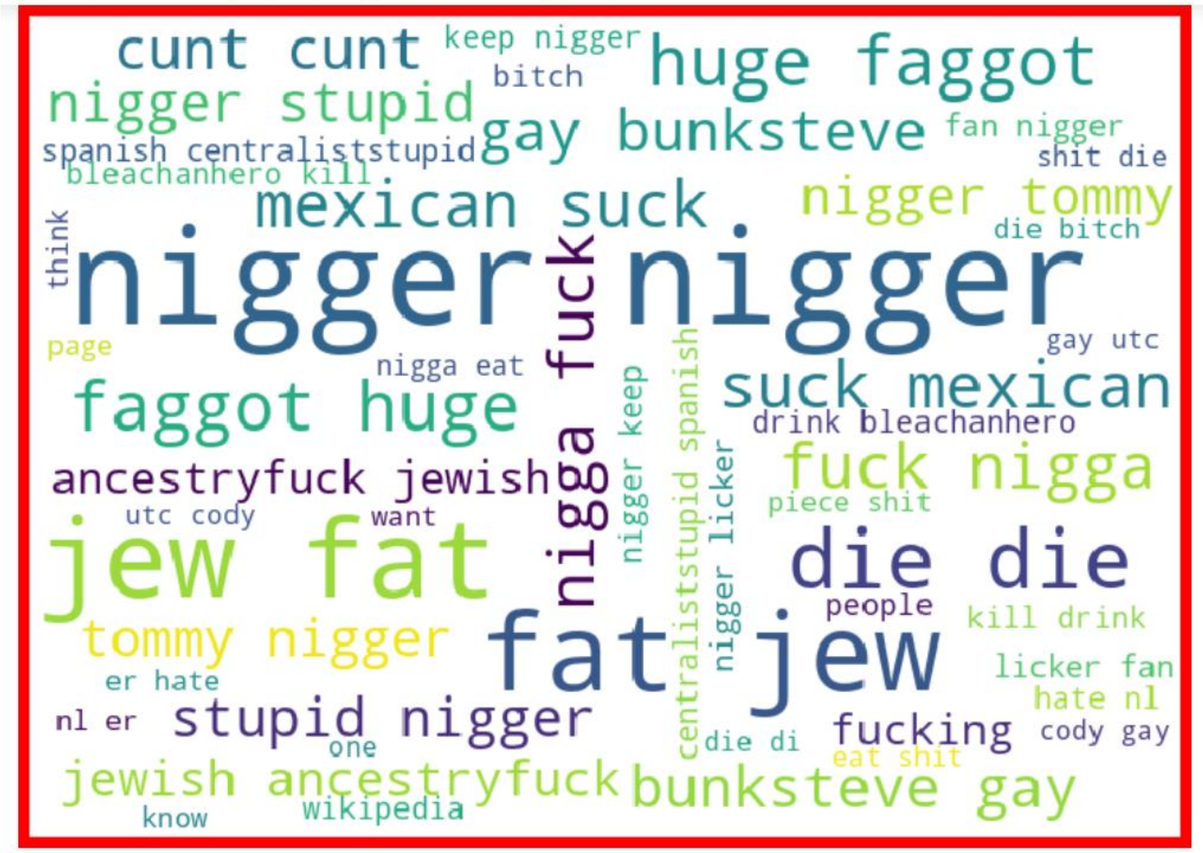
**Abusive**



**Threat**



**Loathe**



**From above we can see, abusive, life threating words are used. Signifying racism , sexism, body shaming comments.**

# Model Deployment

After analyzing the problem statement, the best way to classify fake news into is to use classifications model with the help of Natural Language Processing because classification models take inputs, analyse them and give predicted result.

To evaluate our models we will use :

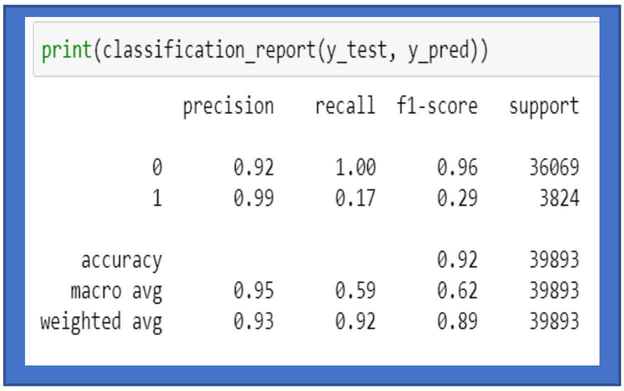
- Confusion matrix
- F1 score
- Recall
- Precision
- Accuracy

- **Precision:** Out of all positives, how many are actually positive.
- **Recall:** Out of all actual positives how many have been predicted as positive.
- **Accuracy:**

**Following Models are used for this study:**

**Malignant comment**

**MultinomialNB**



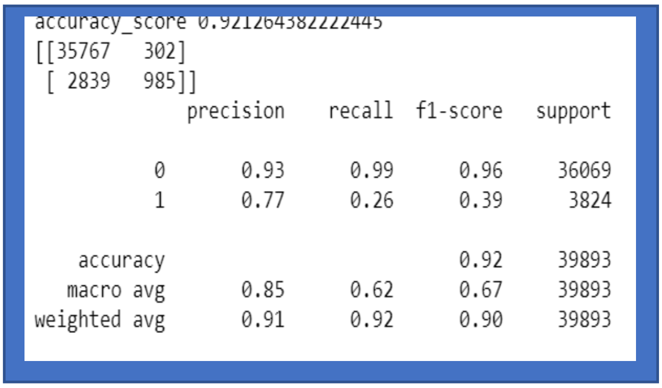
```
print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.92      1.00      0.96     36069
           1       0.99      0.17      0.29      3824

    accuracy                           0.92     39893
   macro avg       0.95      0.59      0.62     39893
weighted avg       0.93      0.92      0.89     39893
```

**Knn**

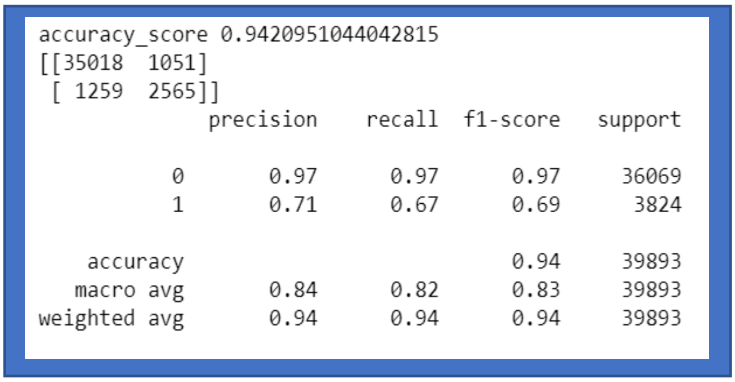


```
accuracy_score 0.921264382222445
[[35767   302]
 [ 2839   985]]
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     36069
           1       0.77      0.26      0.39      3824

    accuracy                           0.92     39893
   macro avg       0.85      0.62      0.67     39893
weighted avg       0.91      0.92      0.90     39893
```

**Decision Tree**



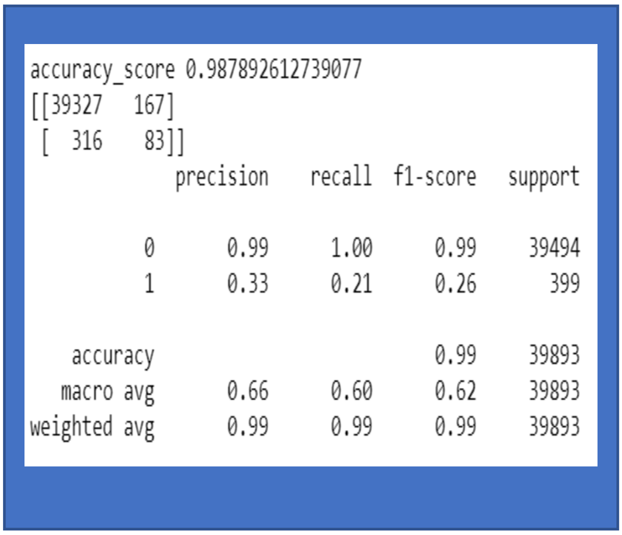
```
accuracy_score 0.9420951044042815
[[35018  1051]
 [ 1259  2565]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     36069
           1       0.71      0.67      0.69      3824

    accuracy                           0.94     39893
   macro avg       0.84      0.82      0.83     39893
weighted avg       0.94      0.94      0.94     39893
```

## Highly malignant

**Knn**

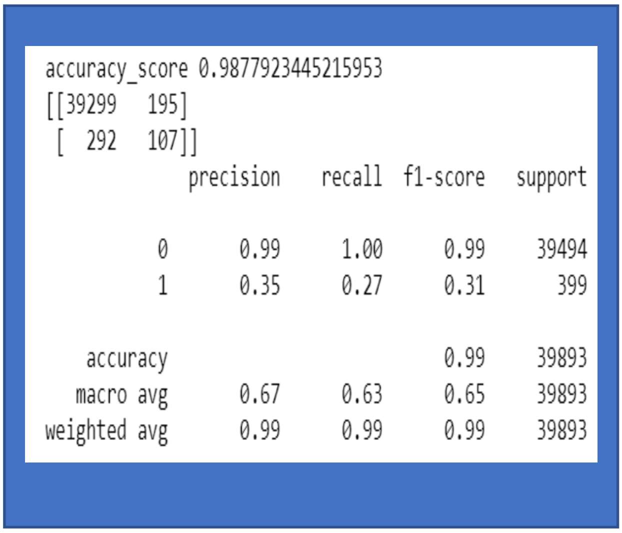


```
accuracy_score 0.987892612739077
[[39327   167]
 [  316    83]]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     39494
           1       0.33      0.21      0.26       399

    accuracy                           0.99     39893
   macro avg       0.66      0.60      0.62     39893
weighted avg       0.99      0.99      0.99     39893
```

**Decision Tree**



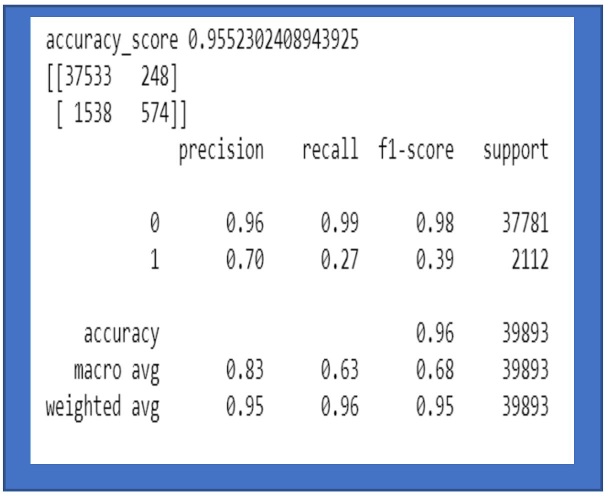
```
accuracy_score 0.9877923445215953
[[39299   195]
 [  292   107]]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     39494
           1       0.35      0.27      0.31       399

    accuracy                           0.99     39893
   macro avg       0.67      0.63      0.65     39893
weighted avg       0.99      0.99      0.99     39893
```

## Rude

**knn**

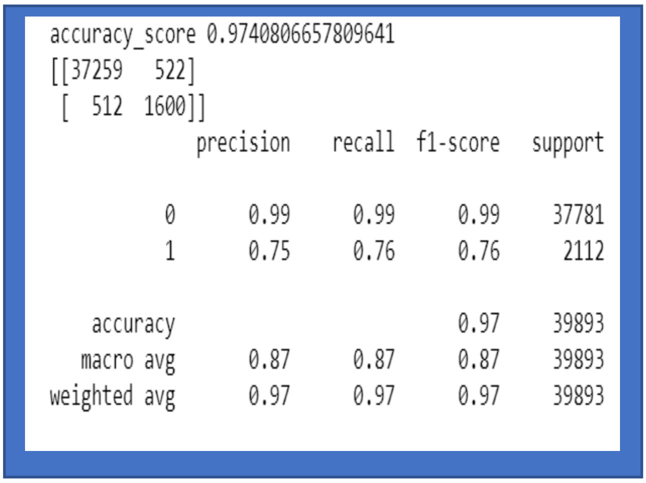


```
accuracy_score 0.9552302408943925
[[37533   248]
 [ 1538   574]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     37781
           1       0.70      0.27      0.39      2112

    accuracy                           0.96     39893
   macro avg       0.83      0.63      0.68     39893
weighted avg       0.95      0.96      0.95     39893
```

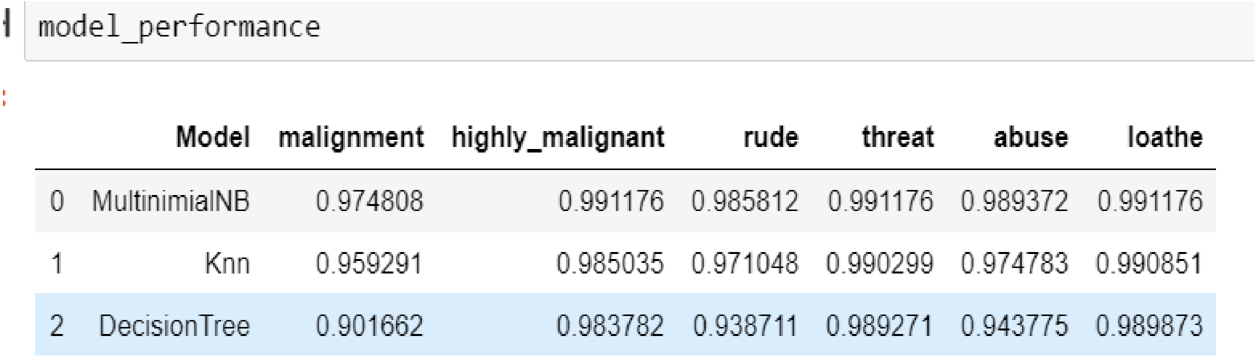
**Decision tree**



```
accuracy_score 0.9740806657809641
[[37259   522]
 [  512  1600]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     37781
           1       0.75      0.76      0.76      2112

    accuracy                           0.97     39893
   macro avg       0.87      0.87      0.87     39893
weighted avg       0.97      0.97      0.97     39893
```



```
model_performance
```

| | Model | malignment | highly_malignant | rude | threat | abuse | loathe |
|---|---|---|---|---|---|---|---|
| 0 | MultinimialNB | 0.974808 | 0.991176 | 0.985812 | 0.991176 | 0.989372 | 0.991176 |
| 1 | Knn | 0.959291 | 0.985035 | 0.971048 | 0.990299 | 0.974783 | 0.990851 |
| 2 | DecisionTree | 0.901662 | 0.983782 | 0.938711 | 0.989271 | 0.943775 | 0.989873 |

# Conclusion

- From the above study it is been observed that highly abusive words like ‘fuck’, ‘suck’, ‘cock’, etc. has been used.
- Also, it is been observed mostly abusive words are based on racism, sexist, body shaming etc.
- Black people are highly abused.
- Multinomial models are the best performing models

In conclusion we can say, malignant comments are widely spread through social media. It not only hurts people’s emotion but also result in act of violence, building model to classify these comments will help to reduce the spread.