**Background of the project**

- Inappropriate emails would demotivates and spoil the positive environment that would lead to more attrition rate and low productivity and Inappropriate emails could be on form of bullying, racism, sexual favoritism and hate in the gender or culture, in today’s world so dominated by email no organization is immune to these hate emails.

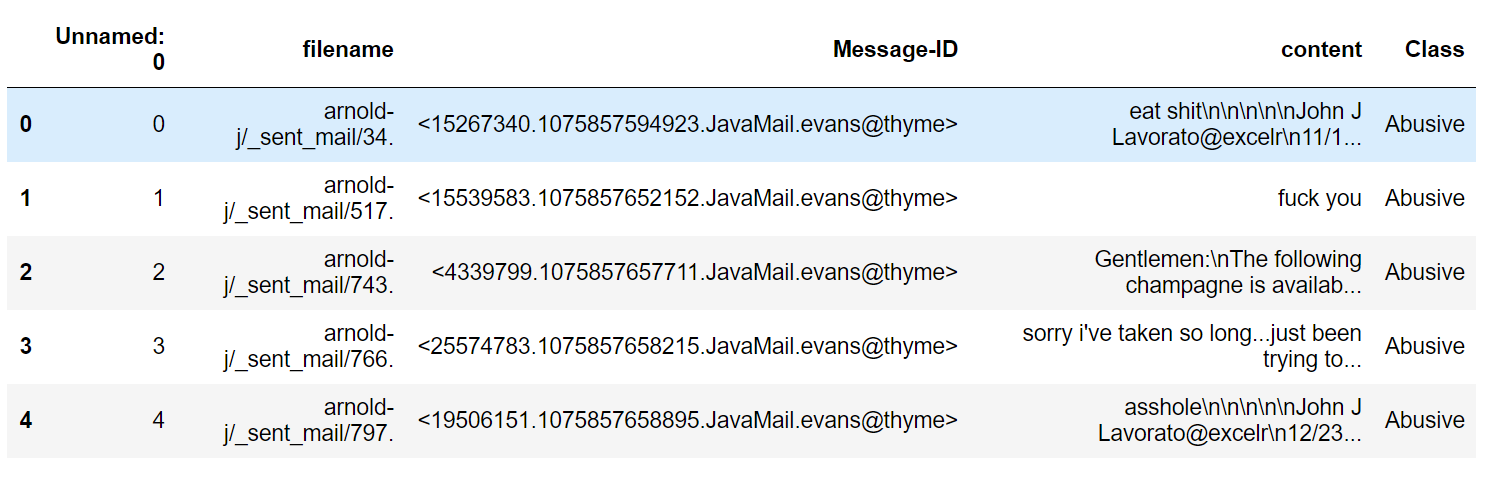
-The goal of the project is to identify such emails in the given day based on the above inappropriate content.

**Problem Statement**

To classify email into abusive and non-abusive

**Dataset Details**

-Data set received in csv file.



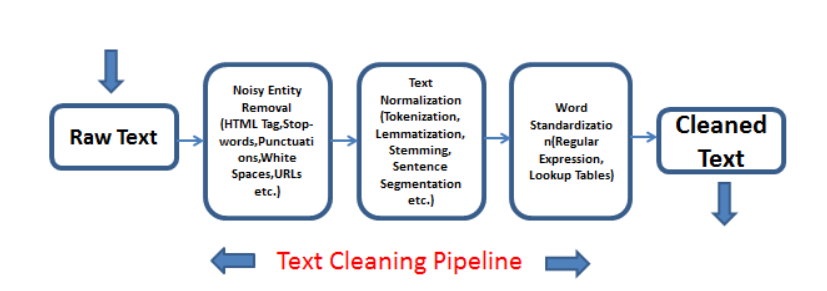
No. of columns =5 (unnamed, filename, Message-ID, content, Class)

No. of Records = *48076*

*Features of Interest: - Independent variable X: - Content*

*Dependent variable Y: -Class*

**For this project we need to do the following task**



* Text Processing
* Text Sequencing
* Model Selection
* Implementation

**Exploratory Data Analysis (EDA)**

1. **Text Processing**

Transforming the raw data into useful format.

The fundamental steps involved in text processing are

1. **Cleaning the raw data**
   1. **Removal of ‘\n’ characters**
   2. **Lowering case**
   3. **Removal of special characters and numbers**
   4. **Removal of Unicode characters**
   5. **Removal of stopwords**
   6. **Removal of hyperlinks**
   7. **Removal of whitespaces**
2. **Text -Normalization – Lemmatization**
3. **Vectorization- Term Frequency Inverse Document Frequency**
4. **Label Encoding –**

Abusive mail is marked as 0

Non -Abusive mail is marked as 1

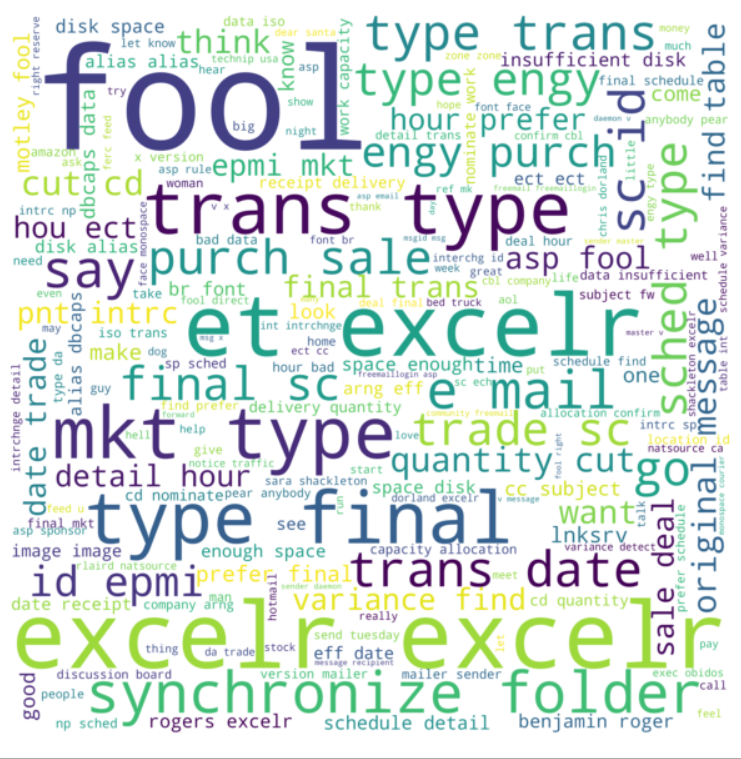
1. **Remove duplicate records**

*25314 are duplicate emails*

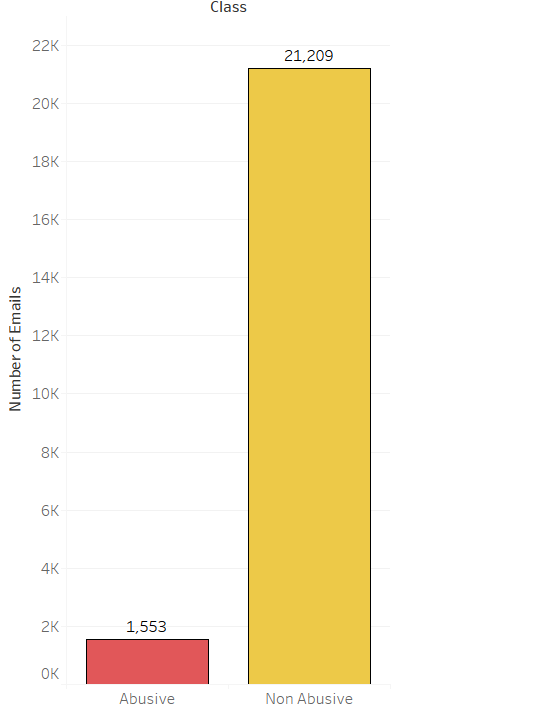
*22762 are unique emails*

1. **Visualization**

**Word cloud**

** **

***Fig:- Word Clouds for abusive emails Fig:- Word Clouds for non-abusive emails***

****

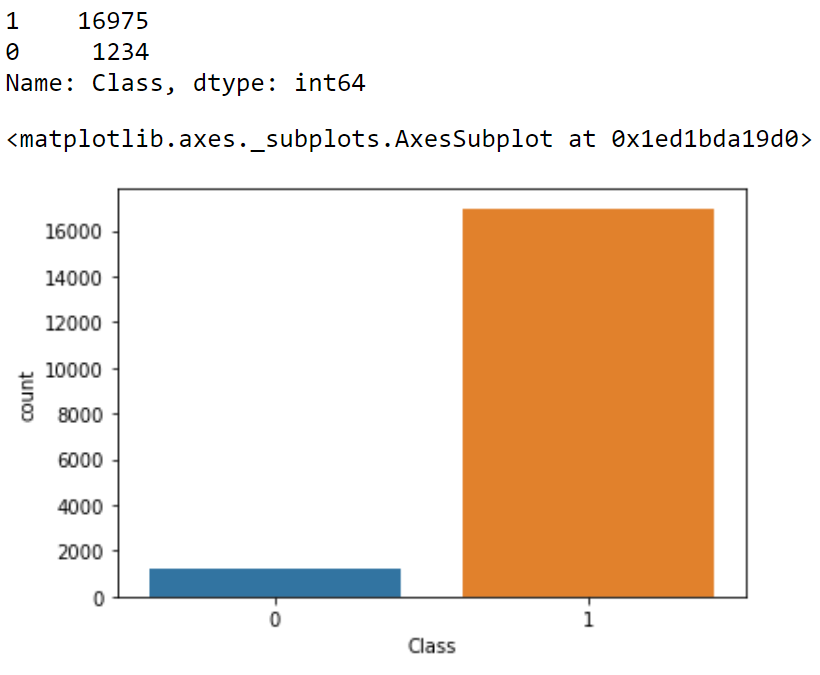
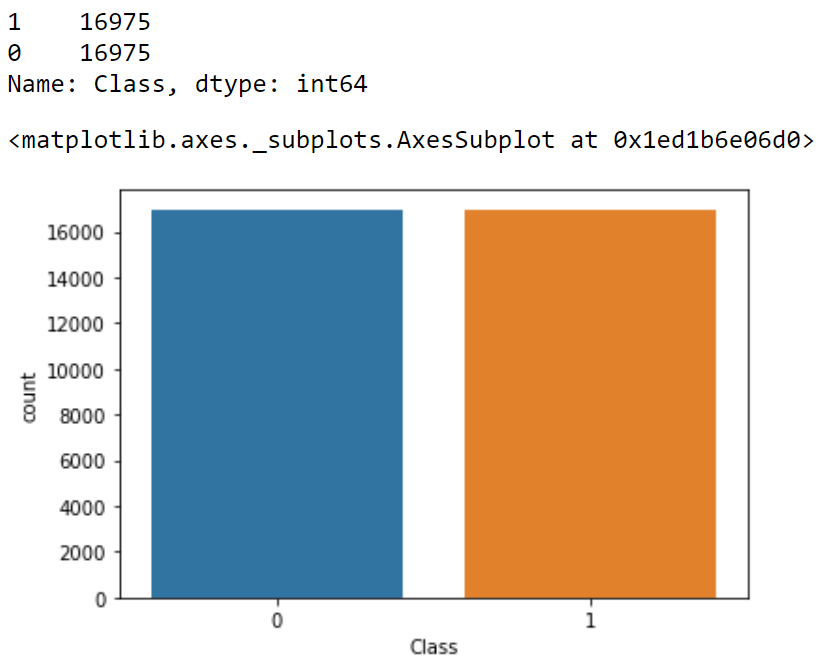
Out of total unique mails only 1553 are Abusive emails and remaining 21,209 are non- abusive mails.

Since there is high differences in the no. of classes, it is a case of imbalanced dataset.

So, we have to balance the dataset.

1. **Balancing Technique**

**SMOTE**

1. **Model Building and Selection**
2. **Implementation**

***Imbalanced Data Balanced Data***

**Model Building**

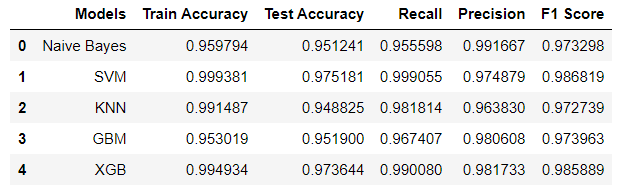
* 1. **Data partitioning**

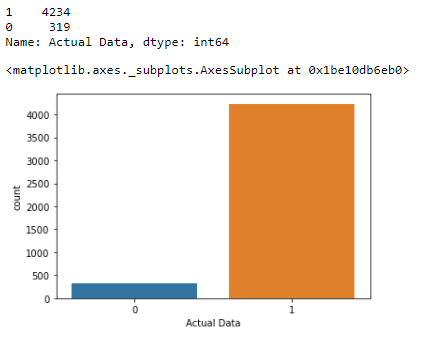
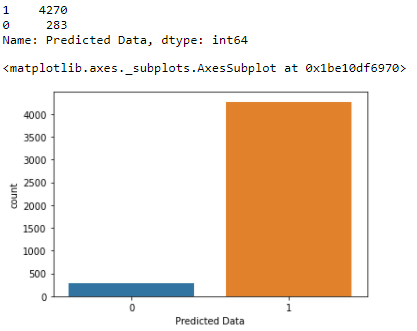
80 % train data

20% test data

* 1. The following models are built
* *Naïve Bayes*
* *Support Vector Machines*
* *K-Nearest Neighbor*
* *Gradient Boosting*
* *XGB*

**Model Evaluation**

****

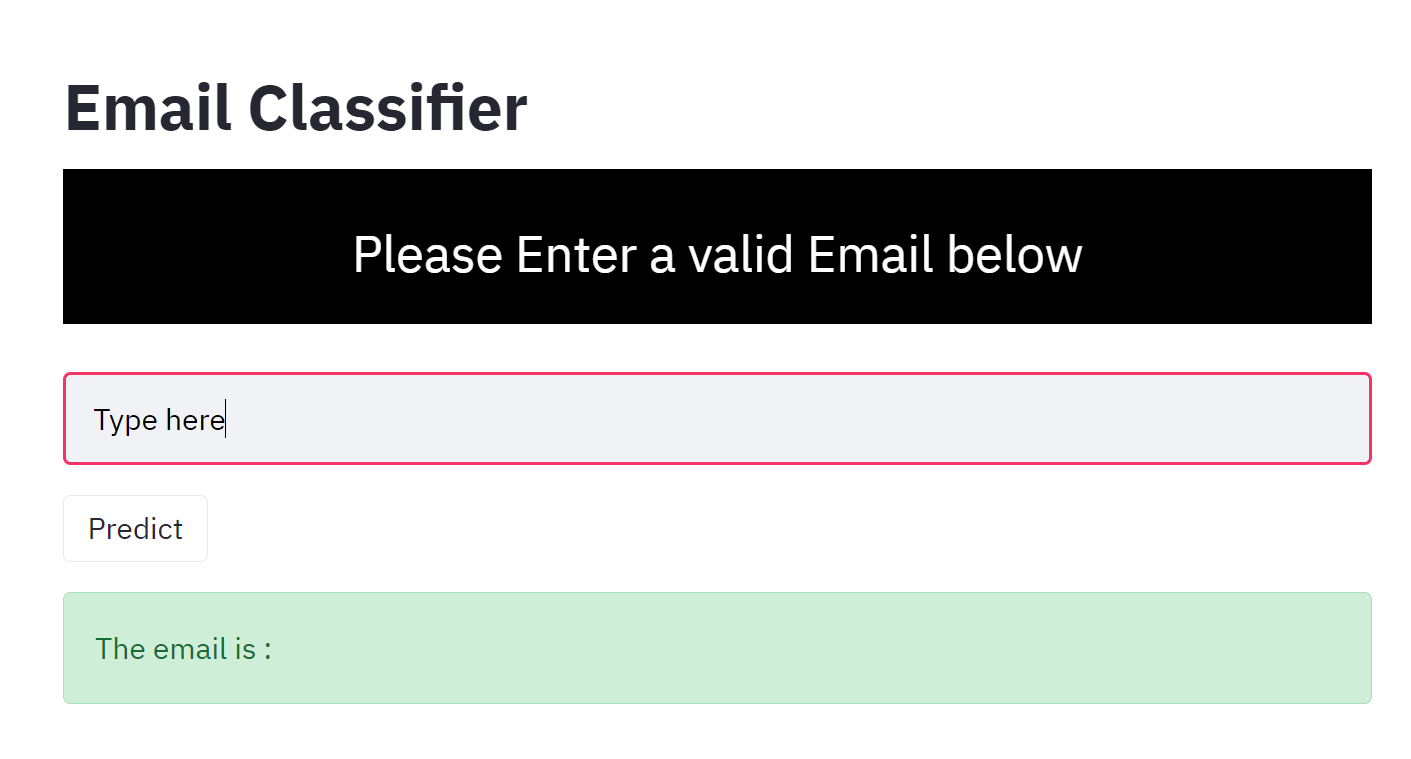
** **

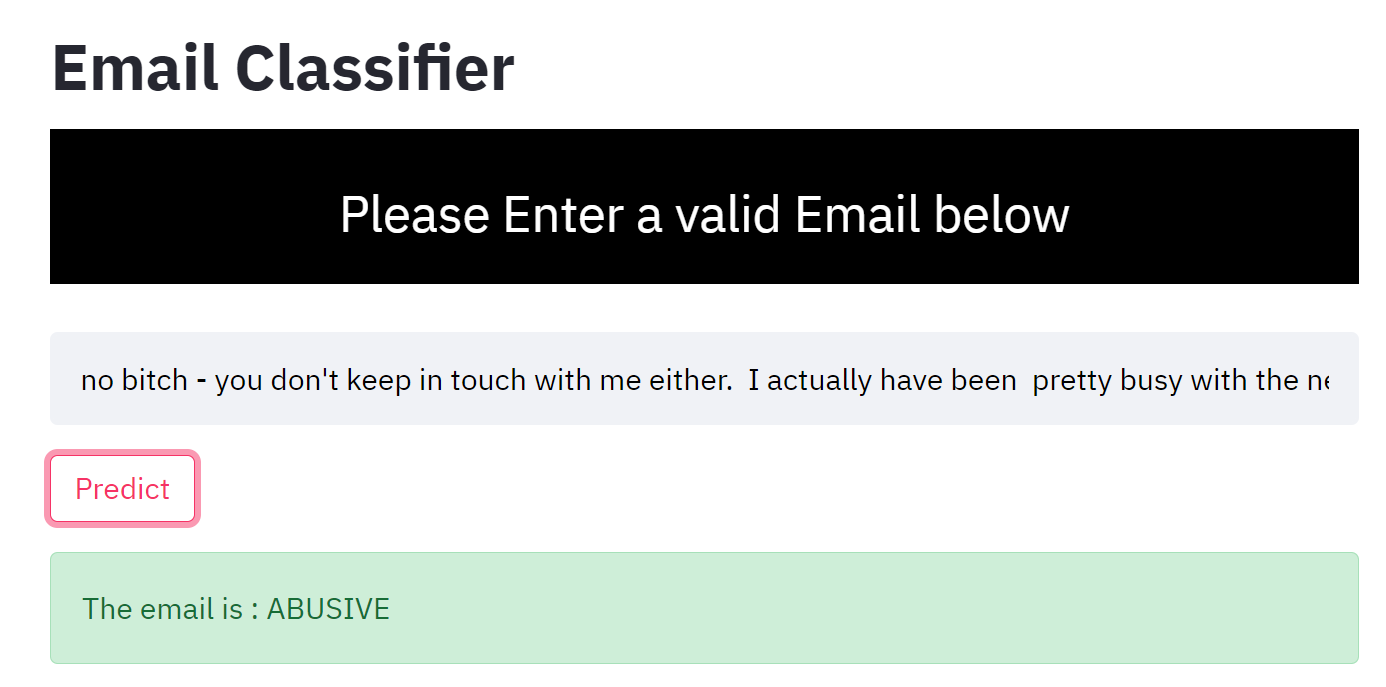
**Confusion Matrix for Model of XGB**

****

**Deployment**

***Used Streamlit for deployment***

****

****