**NLP email1**

**Software - R**

**Business Problem**

**Objective:**

* Inappropriate emails would demotivate and spoil the positive environment that would lead to more attrition rate and low productivity. Inappropriate emails could be in the form of bullying, racism, sexual favourism and hate in the gender or culture. In todays world so dominated by email no organization is immune to these abusive emails.

The goal of the project is to identify such emails in the given day

based on the above inappropriate content.

**Exploratory Data Analysis (EDA) and**

**Feature Engineering**

Dataset: Email1

Rows: 48076

Columns: 5(X, filename, Message.ID, content, Class)

**Exploratory Data Analysis (EDA)** **-Importing dataset into R**

**Cleaning and Processing dataset in R**

* 1. Look for problems in the data such as missing values, impossible values, patterns ,similarities, distributions etc..
  2. Treat all the problems with appropriate techniques.
  3. Apply custom made functions to remove unwanted text and other items.

**Visualizing data in R**

1. Plots

2. Wordcloud

**Model Building**

**Model Results**

Dataset: email1

Rows: 48076

Columns: 5(X, filename, Message.ID, content, Class)

Data Partition details:

Train: 80%

Test: 20%

**Model – Support Vector Machine (SVM)**

Accuracy : 0.9816 Detection Rate : 0.1569

95% CI : (0.9674, 0.9908) Prevalence : 0.1669

No Information Rate : 0.8347 Balanced Accuracy : 0.9687

P-Value [Acc > NIR] : <2e-16 'Positive' Class : Abusive

Kappa : 0.9337

Mcnemar's Test P-Value : 1

Sensitivity : 0.9495

Specificity : 0.9880

Pos Pred Value : 0.9400

Neg Pred Value : 0.9900

Prevalence : 0.1653

**Model - SVM**

**Prediction Abusive Nonabusive**

Abusive 94 6

Nonabusive 5 494

**Model – Naïve Bayes**

**Prediction Abusive Nonabusive**

Abusive8832957

Nonabusive140 10442

    Accuracy : 0.7853           Pos Pred Value : 0.22995

95% CI : (0.7785, 0.7919) Neg Pred Value : 0.98677

    No Information Rate : 0.9291           Prevalence : 0.07093

    P-Value [Acc > NIR] : 1                Detection Rate : 0.06123

    Kappa : 0.2828           Detection Prevalence : 0.26626

    Mcnemar's Test P-Value : <2e-16           Balanced Accuracy : 0.82123

    Sensitivity : 0.86315          Positive' Class : Abusive

    Specificity : 0.77931

**Challenges faced?**

**1**. Dataset was too huge to open and process in 4-12 GB RAM

- Couldn’t even open the csv file as it was a huge file(5 lakhs plus rows)

- Couldn’t even process 10,000 rows for data cleaning

2. Nil to basic understanding of team members about R, techniques/methods

used for data cleaning, model building, shiny app development.

3. Subject Knowledge of team members: nil to basic

4. Availability of team members: zero to minimal

5. Shiny App issues

- Limited understanding of shiny app

- Errors while running the app

**How did you overcome?**

1. Took small amount of data for processing to get accurate results.

2. Adopted an incremental approach to improve efficiency.

3. Pro active approach with mentor in terms of code development, issue

resolution and knowledge sharing**.**