## PROJECT PRESENTATION

**NLP Emails** 

#### Project Team Structure:

Madishetti Rajashekar Project Guide

Bapuram Pallavi
Project Coordinator

Swati Ramniranjan Team member Adwaith Mp Team Member Prajwal Babu Team Member

Rathan Nayaka C Team Member Megha Kulakarni Team Member Shiva Kumar Team Member Dheeraj Team Member

#### Business Objective:

□- 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.

#### Roadmap to Complete the Project



#### Action plan:

SI. No	Stage	Description	Responsibility	Target date	Status	Remarks
1	Frame the problem or understand the business objective	What problem are we trying to solve with NLP? What are the business objectives for using NLP?	All	25 <sup>th</sup> Oct	Closed	
2	Collect the data and understand the data	What data do we need to solve the problem? Where can we get this data? How can we understand the data?	All	27 <sup>th</sup> Oct	Closed	
3	Exploratory data analysis	What are the main characteristics of the data? What are the trends and patterns in the data?	All	31st Nov	Closed	
4	Data cleaning or Text preprocessing	Clean the data to remove noise and inconsistencies. Preprocess the data to put it in a format that can be used by the NLP model.	All	31 <sup>st</sup> Nov	Closed	
5	Vectorization & Model Building	Select the appropriate NLP model for the task. Train the model on the preprocessed data.	All	3 <sup>rd</sup> Nov	Closed	
6	Evaluation of all models and Feedback	Evaluate the performance of the model on a held-out test set. Select the best model based on the evaluation results.	All	14 <sup>th</sup> Nov	Closed	
7	Prediction on new data by deployment	Deploy the model to production and use it to make predictions on new data.	All	14 <sup>th</sup> Nov	Closed	
8	Final presentation	Present the findings and recommendations to the stakeholders.	All	18 <sup>th</sup> Nov	Closed	

#### Data Collection & Overview

#### Name of the data set: emails.csv

Shape of the Data set (48076, 5)

Duplicates Values
0

Null Values/missing values
0

Shape of the Data set (48076, 5)

#### **Data types & Unique Values** Subject:# Column Name D type **U** Values Unnamed: 0 int64 48076 0 filename object 48076 object Message-ID 48076 object 3 23420 content object Class

#### Dataset Cleaning

## Removed Unwanted columns/Features

Removed columns: Index\_No, File\_name, Messiage\_id)

## Removed Duplicates

Shape of the data before (48076, 2) Shape of the data After (24656, 2)

## Final Data set Column D type U Values content object 23420 Class object 2

#### Text Preprocessing

Tokenization

Stop word removal

Transforming to lowercase

Lemmatizing

Removing punctuations

Removing URLs

Removing numeric values

Removing unwanted spaces

Removing special characters

Removing unwanted spaces

Dropping Null values

Removing duplicate values

**Shape of data(22811, 2)** 

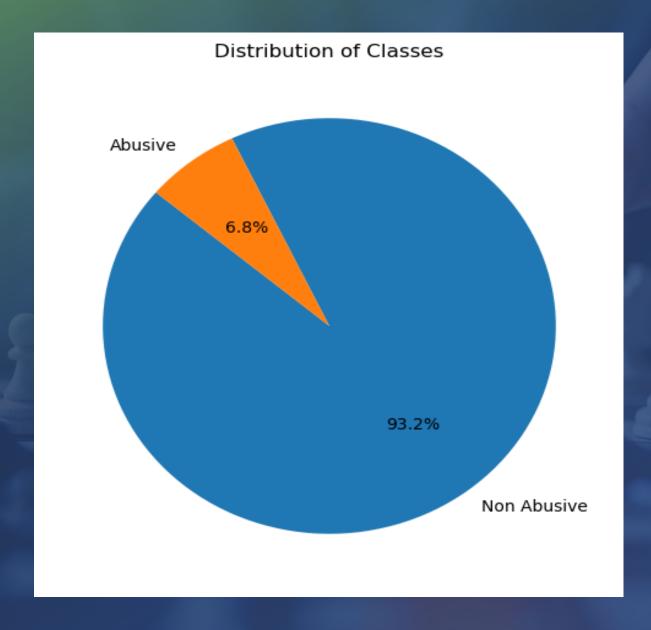
#### EDA - Calculated statistics

Average sentence length 1182.187

Number of unique words

Most frequent words

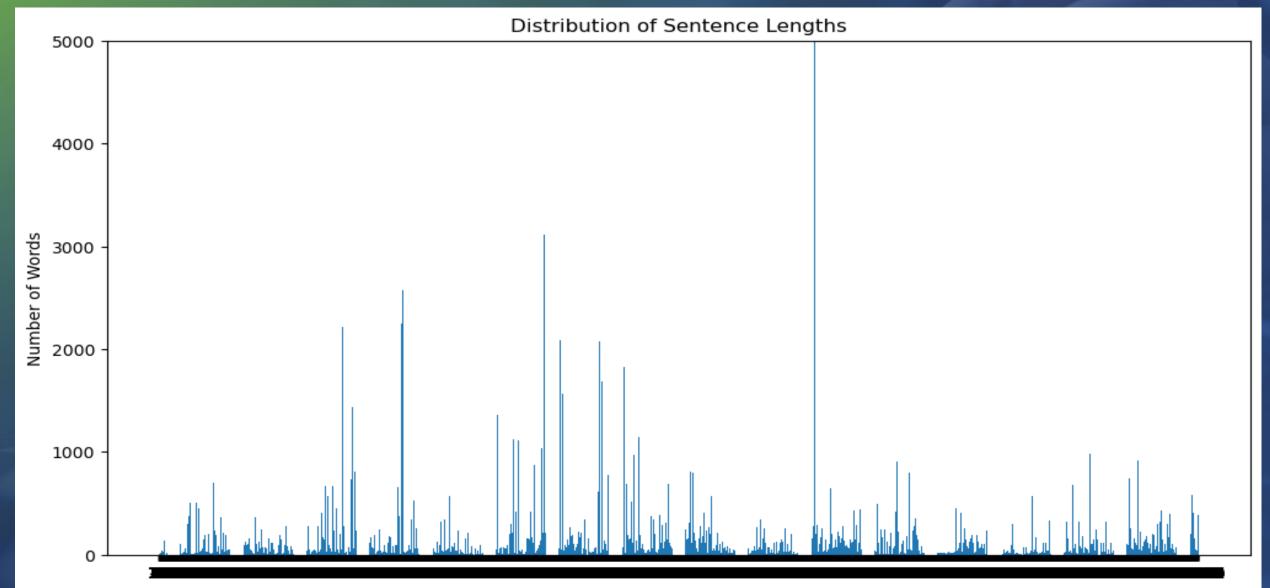
#### EDA - Data Visualization- Class distribution



#### EDA - Data Visualization- World Cloud



#### EDA - Data Visualization- Mail wise No.of words



#### EDA

#### **Sentiment Analysis**

Assign a sentiment score to each email.

#### **Topic Modelling**

Identify key topics within the emails.

#### Part-of-Speech (POS) Tagging

Categorize words into their respective parts of speech.

#### Vectorization

#### Independent variable

Dependent variable

- TF-IDF (term frequency-inverse document frequency) vectorization to convert final text tokens data into numerical features
- Shape of the data after vectorization (22811, 117893)
- Encoded the target variable(classes) using label encoding
- Abusive -0
- Non-abusive 1

#### Model Building & Evaluation

Split the test data and

Cross fold method (CV=5)

**Model Evaluation Aaccuracy Scores** 

**Models Selection** 

Model prediction on test data

Model training on Train data

#### Model Building & Evaluation

S.No	Model Name	Accuracy Score
1	XG Boost Classifier	96.40%
2	Gradient Boosting Classifier	95.93%
3	Support vector Classifier	95.80%
4	Logistic regression	95.44%
5	AdaBoost Classifier	95.29%
6	Random Forest Classifier	95.10%
7	Decision Tree Classifier	95.08%
8	K-Nearest neighbours	93.91%
9	Multinomial NB	93.24%

- •Presented accuracy scores for each model.
- •XG Boost Classifier is the best-performing model for deployment.

#### Model Deployment with Streamlit

Import pre-trained TF-IDF vectorizer and XGBoost model on Streamlit local host



Create interactive Streamlit web app on Streamlit local host



Load pre-trained TF-IDF vectorizer and selected model



Implement text preprocessing functions in streamlit app



Display prediction results



Deploy XGBoost model with highest accuracy



Create user interface for email content input

#### Model Deployment with Streamlit

#### **Email Classification**

Enter an email to check if it's appropriate or not. **Email Content:** awesome Check Email **Non-Abusive** This email is appropriate or Non-Abusive.

#### **Email Classification**

Enter an email to check if it's appropriate or not.

Email Content:

shit

Check Email

This email is inappropriate or Abusive.

**Abusive** 

Sample prediction results snapshots from localhost

#### Challenges faced during a project:

□The dataset may be imbalanced, with a significantly higher number of normal emails compared to inappropriate ones.

□ Faced challenges in text preprocessing to effectively capture and represent inappropriate language nuances.

□ Faced computational resource challenges, particularly with larger datasets and resource-intensive models like Random Forest or Gradient Boosting.

# Thank you