# Business Problem:

▪ 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 favourism 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.

Objective:

The main objective of the project is to develop an accurate predictive Text Classification Model through the Python language, which determines whether an email-message is Abusive or Non-Abusive.

Project Architecture / Project Flow:
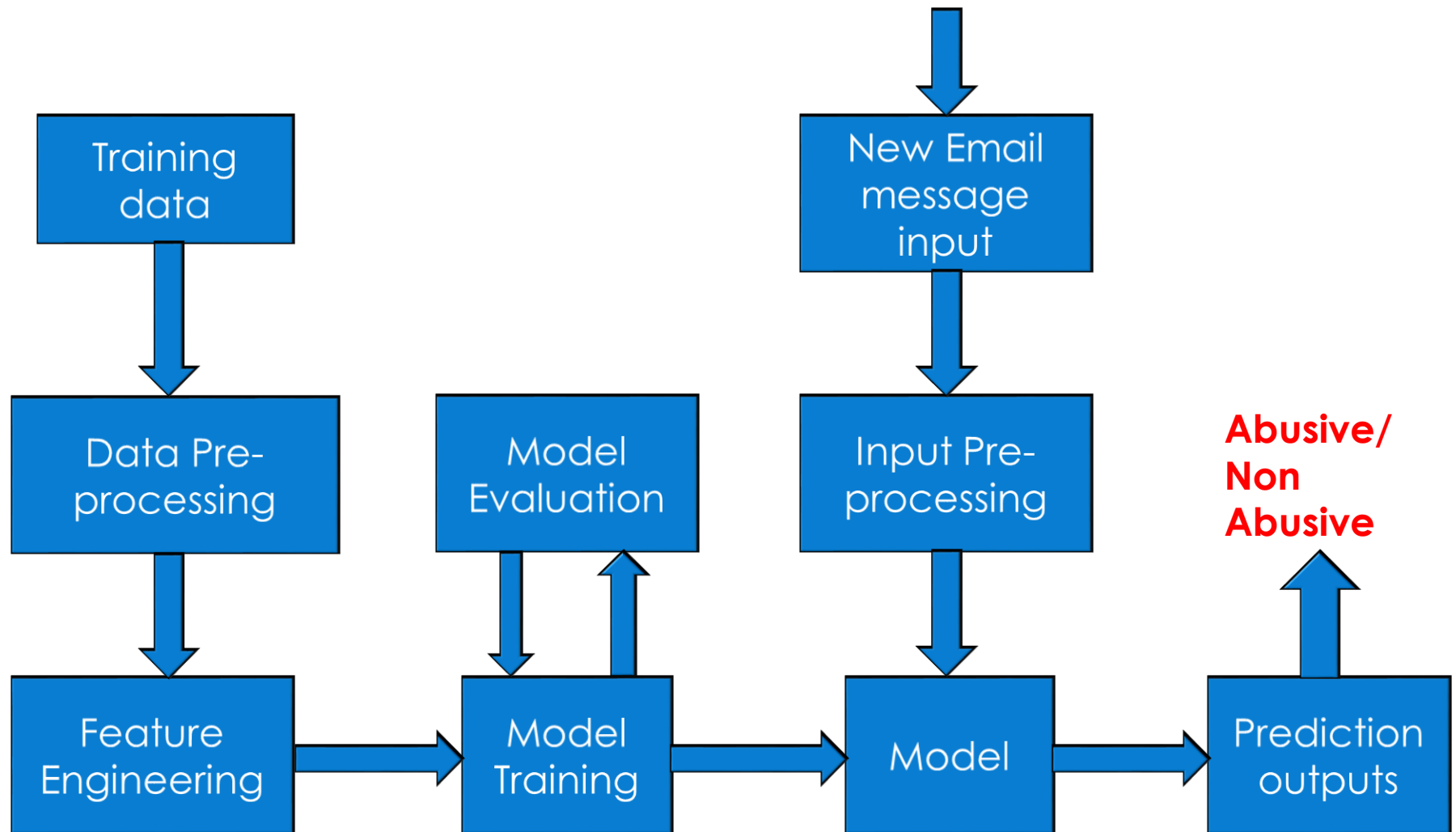
Data set details:

Email messages data set with labeled data:-

It contains,

▪ Total **Rows = 48076** & **columns = 5**

▪ columns = ('Unnamed: 0', 'filename', 'Message-ID', 'content', 'Class')

▪ Drop unwanted columns which are ('Unnamed: 0', 'filename', 'Message-ID')

▪ final data set contains only 2 columns namely **('content' & 'Class')**

▪ **'content'** contains **text email messages** which are 'features'.

▪ **'Class'** contains **'Abusive' & 'Non Abusive'** categories which are labels.

▪ count values for target columns = Non Abusive : 44666

                                                          Abusive : 3410


▪ There are '0' missing(null) value in all columns.

	Unnamed: 0	filename	Message-ID	content	Class
0	0	arnold-j/_sent_mail/34.	<15267340.1075857594923.JavaMail.evans@thyme>	eat shit\n\n\n\n\nJohn J Lavorato@excelr\n11/1...	Abusive
1	1	arnold-j/_sent_mail/517.	<15539583.1075857652152.JavaMail.evans@thyme>	fuck you	Abusive
2	2	arnold-j/_sent_mail/743.	<4339799.1075857657711.JavaMail.evans@thyme>	Gentlemen:\nThe following champagne is availab...	Abusive
3	3	arnold-j/_sent_mail/766.	<25574783.1075857658215.JavaMail.evans@thyme>	sorry i've taken so long...just been trying to...	Abusive
4	4	arnold-j/_sent_mail/797.	<19506151.1075857658895.JavaMail.evans@thyme>	asshole\n\n\n\n\nJohn J Lavorato@excelr\n12/23...	Abusive

Exploratory Data Analysis (EDA) & Feature Engineering:

▪ Drop unwanted first 3 columns.
❑ Details of Class & content column…

	Class	Abusive	Non Abusive
content	count	3410	44666
	unique	1642	23014
	top	\n\n -----Original Message-----\nFrom: \tCusto...	Ken Lay and Jeff Skilling were interviewed on ...
	freq	11	19

❑ Distribution of Abusive & Non-Abusive Data:-
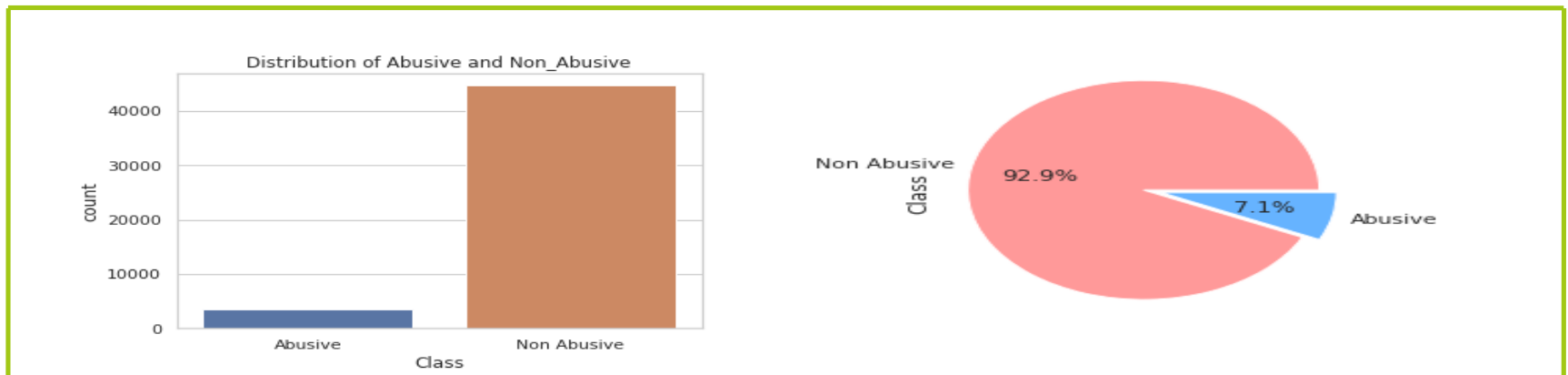
▪From above two figures it is clear that Abusive & Non-Abusive data has **Class imbalance** problem.
▪ so we apply the sampling technique (over-sampling)

EDA & Feature Engineering:-

☐Most frequent words in Abusive & Non-Abusive class data:

▪Majority of most frequent words in both classes are stop words such as 'to', 'the', 'a' etc.
▪ so remove these stop words by putting them into 'Word Cloud'.

☐After removing ('english') stop words... **Word Cloud**:-

Data Pre-processing:-

❑ **Text Cleaning:**
To clean the data, explore some common Techniques of NLP…
For that define the 'clean_text' function…then,
   1) Lower the text & remove spaces
   2) Remove the text between square brackets
   3) Remove the punctuations
   4) Remove html strips
   5) Tokenize the text
   6) Remove stop words (unwanted words like…excelr, hou,)
   7) Apply Porter Stemmer (stemming)
Apply this 'clean_text' function on 'content' column(which is feature).

❑ **Features & Labels:**
▪ X = email('content')   …features
▪ y = email('label')   …labels

❑ **Feature Extraction:**
▪ Used **Tfidf Vectorizer** & take max_features = 10000
▪ then fit the data by 'fit_transform'

Model Building:-

▪ First of all, split the data into train & test by 70/30 ratio.

▪We building our model on 6 different Machine Learning algorithms.
▪They are,
        1) Random Forest,
        2) Naive Bayes,
        3) Support Vector Machine,
        4) Decision Tree,
        5) Extra Tree classifier,
        6) Bagging Classifier   (ensemble classifier)
The purpose behind this is deciding which perform the best.

▪ Above all algorithms gives better accuracy(>95%) but we choose
  Random Forest from them because it gives high Accuracy, better TPR, FPR,
TNR, FNR, recall, & f1-score as compared to others.
▪It has the power of handle large data sets with higher dimensionality.
▪ It has methods for balancing errors in data sets where classes are
imbalanced.
▪ It helps to overcome the problem of over fitting.

Template for Model results presentation:-

Model – Random Forest

Data set details:
- Total Rows = 48076, columns=5
- In Tfidf take max_features = 10000

Data Partition details:
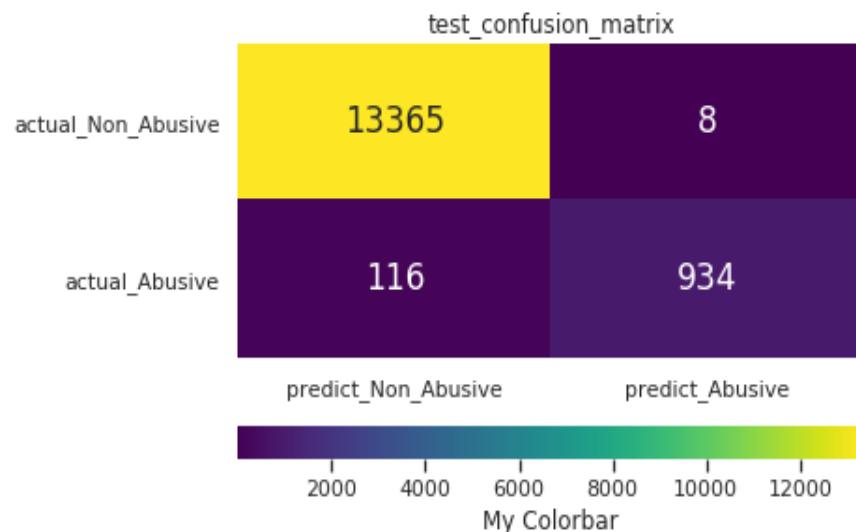- Split the data into train & test by **70:30** ratio.

Algorithm: Random Forest Classifier

Algorithm details and configuration:
bootstrap=True,
criterion='gini',
n_estimators=100,
random state=252
max_features='auto',
max_depth=None,   min_samples_split=2,
max_leaf_nodes=None, n_jobs=None.

Results:-   Accuracy: **99.11%**
Confusion matrix:

test_confusion_matrix

	predict_Non_Abusive	predict_Abusive
actual_Non_Abusive	13365	8
actual_Abusive	116	934

Classification Report:

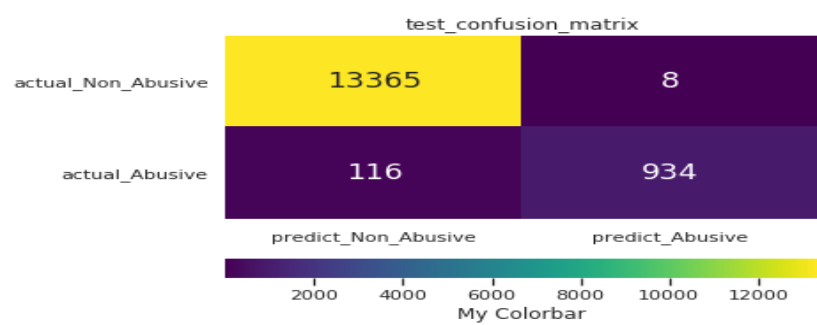|	precision	recall	f1-score	support
0	0.99	1.00	1.00	13373
1	0.99	0.89	0.94	1050
accuracy			0.99	14423
macro avg	0.99	0.94	0.97	14423
weighted avg	0.99	0.99	0.99	14423

Model Performance:-

❑Check whether Model is **over fit** or not by using cross-validation dataset:-

Training Data	*Cross-validation(test) Data*

✓**Accuracy = 99.11%**   **98.65%**

✓**Confusion Matrix:**

test_confusion_matrix

	predict_Non_Abusive	predict_Abusive
actual_Non_Abusive	13365	8
actual_Abusive	116	934

2000 4000 6000 8000 10000 12000
My Colorbar

test_confusion_matrix

	predict_Non_Abusive	predict_Abusive
actual_Non_Abusive	9332	14
actual_Abusive	122	628

2000 4000 6000 8000
My Colorbar

✓**Classification Report:**

	precision	recall	f1-score	support
0	0.99	1.00	1.00	13373
1	0.99	0.89	0.94	1050
accuracy			0.99	14423
macro avg	0.99	0.94	0.97	14423
weighted avg	0.99	0.99	0.99	14423

	precision	recall	f1-score	support
0	0.99	1.00	0.99	9346
1	0.98	0.84	0.90	750
accuracy			0.99	10096
macro avg	0.98	0.92	0.95	10096
weighted avg	0.99	0.99	0.99	10096

▪By observing both train & cross-validation dataset, it is clear that Accuracy of both almost same(no drastic difference).

▪Also precision, recall & f1-score observe to be close.
    so it's clear that our **model is best fit** & there is no over fit as well as under fit.

Model Deployment using Flask Framework:-

▪With help of Flask web framework we build the web application.
▪we upload test dataset in '.csv' format.

❑ we did it for both **text box** & **csv file** to obtain the predictions:

choose you have single input or .CSV file

◉ **Single string input**

enter your input

I am stuck in traffic now

Text box ➡

.csv file ➡ ○ **.csv file**

[Predict] [Reset]

▪At a time it gives only one output either from 'text box' or '.csv file'
▪ In 'text box' put whatever string input it gives **Abusive** or **Non Abusive** as an **output**.
▪For '.csv file' put an test data csv file & it gives the prediction of whole observations as an **Abusive** or **Non Abusive** in '**csv file' format.**

Results/Predictions obtain from deployment:-

❑ Output of single string:

Abusive Prediction

← → C ⓘ 127.0.0.1:5000/predict

ML App

Abusive Detector

↓

predicted

	content	predicted
0	cousin told chew ass wa pretti embarass realli nice guy	Abusive

Non Abusive Prediction

← → C ⓘ 127.0.0.1:5000/predict

ML App

Abusive Detector

↓

predicted

	content	predicted
0	yeah\r\ni will be there in 5 minutes\r\nsurely	Non Abusive

❑ Output of csv file:

Select the test file

choose you have single input or .CSV file

◯ Single string input

◉ .csv file

  choose your csv file
  [Choose File] test.csv

[Predict] [Reset]

Results →

Predictions of whole observations

A1   fx No.

	A	B	C
1	No.	content	Predicted
2	0	travel plan final plea help us plan provid itinerari in	Non Abusive
3	1	dear mr bass thank electron check request submit r	Non Abusive
4	2	greg piper salli beck mark picker review document p	Non Abusive
5	3	comput commiss day move new hou therefor abl se	Non Abusive
6	4	diana info request per conver earlier plea tri arrang	Non Abusive
7	5	thi great want use upcom credit confer thank rick	Non Abusive
8	6	see new sap code old e work order invoic tj forward	Non Abusive
9	7	ramiro last name	Non Abusive
10	8	elect integr internet activist keep democrat leaders	Non Abusive
11	9	sound great keep good work power god bobett riner	Non Abusive
12	10	calendar entri event descript date durat day chairpe	Non Abusive
13	11	hey yvett made reserv nino thank eric yvett g eric br	Non Abusive
14	12	forward phillip k allenhouect pm tim heizenrad phil	Non Abusive
15	13	hey hope weekend wa good think still plan come ha	Non Abusive
16	14	attatch com ed entergi vol curv look histor daili price	Non Abusive
17	15	fyi might back offic issu thi forward eric basshouect	Non Abusive
18	16	websupport john arnold cc subject portfolio jennife	Non Abusive

Challenges faced?

- problem faced during importing the big size csv file(big size data)
- Imbalanced class data…
- During cleaning, what NLP technique should be preferred stemming or lemmatizing.
- Proper Machine Learning Algorithm selection…
- At the time of deployment of model in flask, it gives errors continuously like…value error, memory error etc.

How did you overcome?

- ✓ first did it on some small samples & then perform in Google-colab.
- ✓ for imbalance class data apply sampling techniques & also cross verify it by checking TPR, FPR, TNR & FNR ratios.
- ✓ By performing both stemming & Lemmatizing, which gives better results we choose stemming.
- ✓ Most of the Algorithms gives accuracy>95%, but accuracy is not only the criteria to define a model is best fit or not. So check over fit, confusion matrix, recall, f1-score & mathematical structure behind the algorithm.
- ✓ searched from Google, kaggle & other resources and apply same on data.

Thank you