Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

BE-ETRX **UID:**2019110039 **Sub-Minor ML**
NAME: Devansh Palliyath

Exp 4

Aim: Apply Naive bias on an NLP dataset.

Objective:

1. Loading the dataset
2. Performing EDA
3. Implementing Naive Bias algorithm on the NLP dataset.
4. Training the dataset
5. Evaluating the model generated.
6. Studying accuracy, recall, precision etc.

Dataset:
https://www.kaggle.com/datasets/datatattle/email-classification-nlp

Code:

```
#For general purpose

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

```
#For data preprocessing

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.model_selection import train_test_split

#For model creation and training

from sklearn.naive_bayes import MultinomialNB

#For model evaluation

from sklearn.metrics import classification_report, confusion_matrix

file_train = '/content/SMS_train.csv'

file_test = '/content/SMS_test.csv'



df_train = pd.read_csv(file_train, encoding = 'cp1252')

df_test = pd.read_csv(file_test, encoding = 'cp1252')

print(df_train.head())

print(df_train.shape)
```

```
   S. No.                                       Message_body     Label
0       1                       Rofl. Its true to its name  Non-Spam
1       2  The guy did some bitching but I acted like i'd...  Non-Spam
2       3  Pity, * was in mood for that. So...any other s...  Non-Spam
3       4              Will ü b going to esplanade fr home?  Non-Spam
4       5  This is the 2nd time we have tried 2 contact u...      Spam
(957, 3)
```

```
print(df_train.head())

print(df_train.shape)
```

```
   S. No.                                        Message_body     Label
0       1                          Rofl. Its true to its name  Non-Spam
1       2  The guy did some bitching but I acted like i'd...  Non-Spam
2       3  Pity, * was in mood for that. So...any other s...  Non-Spam
3       4              Will ü b going to esplanade fr home?  Non-Spam
4       5  This is the 2nd time we have tried 2 contact u...      Spam
(957, 3)
```

```
df_train['y'] = pd.Categorical(df_train['Label']).codes

print(df_train.head())

print(df_train.shape)
```

```
   S. No.                                        Message_body     Label  y
0       1                          Rofl. Its true to its name  Non-Spam  0
1       2  The guy did some bitching but I acted like i'd...  Non-Spam  0
2       3  Pity, * was in mood for that. So...any other s...  Non-Spam  0
3       4              Will ü b going to esplanade fr home?  Non-Spam  0
4       5  This is the 2nd time we have tried 2 contact u...      Spam  1
(957, 4)
```

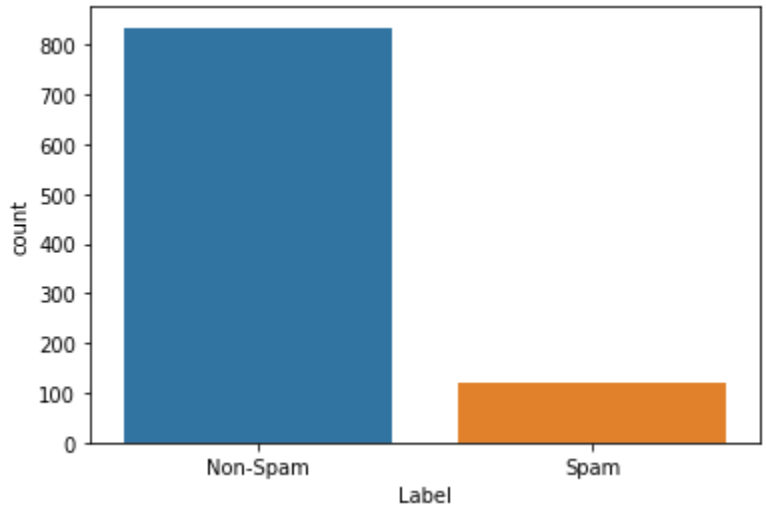
```
df_test['y'] = pd.Categorical(df_test['Label']).codes

print(df_test.head())

print(df_test.shape)
```

```
   S. No.                                        Message_body Label  y
0       1  UpgrdCentre Orange customer, you may now claim...  Spam  1
1       2  Loan for any purpose £500 - £75,000. Homeowner...  Spam  1
2       3  Congrats! Nokia 3650 video camera phone is you...  Spam  1
3       4  URGENT! Your Mobile number has been awarded wi...  Spam  1
4       5  Someone has contacted our dating service and e...  Spam  1
(125, 4)
```

```
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   S. No.        125 non-null    int64
 1   Message_body  125 non-null    object
 2   Label         125 non-null    object
 3   y             125 non-null    int8
dtypes: int64(1), int8(1), object(2)
memory usage: 3.2+ KB
```

```
sns.countplot(df_train['Label'])
```




```
df = pd.concat([df_train,df_test], ignore_index=False, axis=0)

df.shape

df.shape
```

```
(1082, 4)
```

```
y = df['y'].values

y.shape

vectorizer = CountVectorizer()

spamham_countVector = vectorizer.fit_transform(df['Message_body'])
```

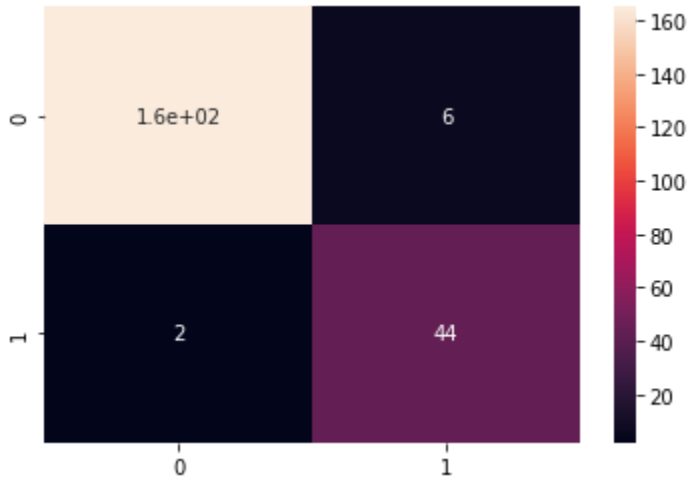
```
spamham_countVector.shape

X_train, X_test, y_train, y_test = train_test_split(spamham_countVector,
y, test_size = 0.2)

print(X_train.shape)

print(y_train.shape)

print(X_test.shape)

print(y_test.shape)
```

```
(865, 3527)
(865,)
(217, 3527)
(217,)
```

```
# Train the model using naive bias

NB_classifier = MultinomialNB()

NB_classifier.fit(X_train,y_train)
```

```
MultinomialNB()
```

```
# Evaluate the model
y_pred = NB_classifier.predict(X_test)

cm = confusion_matrix(y_test, y_pred)

sns.heatmap(cm, annot=True)
```



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

```
              precision    recall  f1-score   support

           0       0.99      0.96      0.98       171
           1       0.88      0.96      0.92        46

    accuracy                           0.96       217
   macro avg       0.93      0.96      0.95       217
weighted avg       0.97      0.96      0.96       217
```

Conclusion:

- The Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.
- For this dataset we used NLP and naive bayes, for NLP we require multinomial naive bayes.