AI FOR CYBER SECURITY ASSIGNMENT-1 REPORT

MUHAMMED FAVAS – 142102007

TASK-1: Naive Classifier

Preprocessing applied:

- Extracted the mail body and subject
- Converted mail ids to a string ‘mailaddress’
- Removed non-alphanumeric characters
- Tokenized each text
- Converted all characters to the lowercase
- Removed stopwords
- Removed tokens with length less than 3.
- Applied stemming
- Created features(X) as tokenized words and labels(y) as 0 for ham mails and 1 for spam mails
- Split whole training data to train and test samples in the ratio of 7:3

Training:

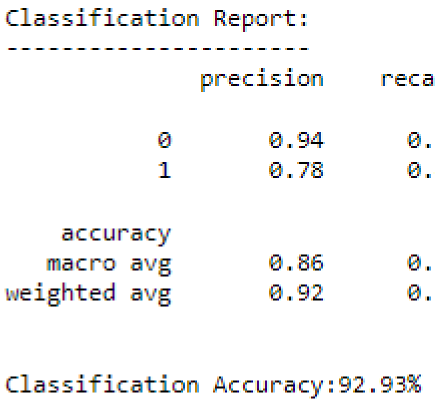
Found set of blacklist tokens as follow

- Found set of unique words from all the spam mails as positive words
- Found set of unique words from all the ham mails as negative words
- Subtracted set negative words from set of positive words. Which gives blacklist

Prediction:

- Found set of unique word in the test sample
- If the size of intersection this set with blacklist set greater than or equal to 1, test sample classified as spam, else ham.

Results:

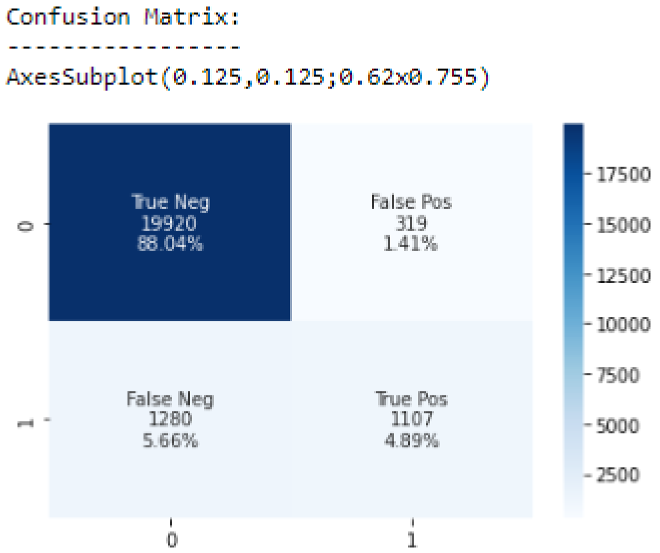


```
Classification Report:
----------------------
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     20239
           1       0.78      0.46      0.58      2387

    accuracy                           0.93     22626
   macro avg       0.86      0.72      0.77     22626
weighted avg       0.92      0.93      0.92     22626


Classification Accuracy:92.93%
```



```
Confusion Matrix:
-----------------
AxesSubplot(0.125,0.125;0.62x0.755)
```

TASK-2: NAIVE BAYES CLASSIFIER

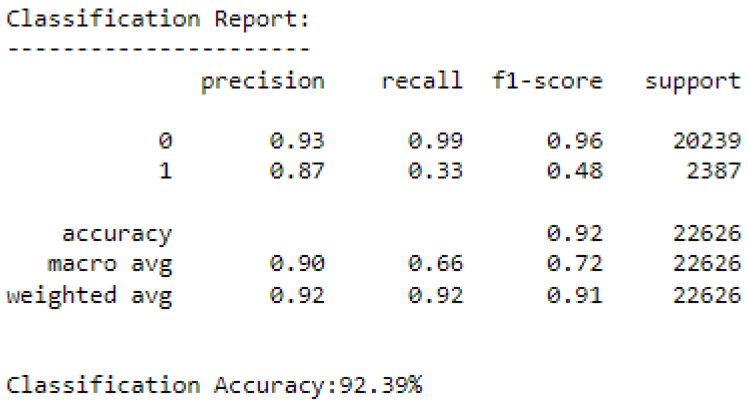
Preprocessing:

- Joined all tokenized words used in task-1 to texts for each mails, which gives the corpus.
- Vectorized each mails with TFIDF vectorizer from scikit library, which gives feature vectors corresponding to each mail.
- Split training samples to train and test samples in the ratio of 7:3

Training and Prediction:

Using Multinomial Naïve Bayes Classifier library from scikit learn trained the classifier with traning samples and predicted the output labels from test samples

Results:



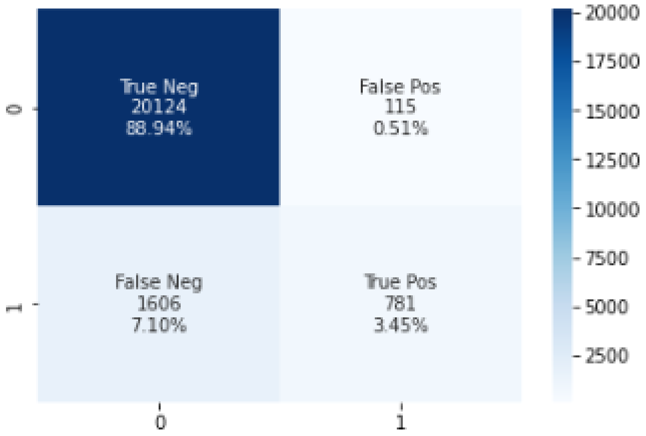
```
Classification Report:
----------------------
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     20239
           1       0.87      0.33      0.48      2387

    accuracy                           0.92     22626
   macro avg       0.90      0.66      0.72     22626
weighted avg       0.92      0.92      0.91     22626


Classification Accuracy:92.39%
```

```
Confusion Matrix:
-----------------
AxesSubplot(0.125,0.125;0.62x0.755)
```


True Neg
False Pos
False Neg
True Pos