The Project Report of Email Spam Detection

This Email Spam detection file contains the 2893 rows and 3 columns.
The missing value in this file is 62 which is in subject data.

Data Preprocessing

It contains the three columns name *Subject*, *Message* and *label.*
The shape of this file is (2893, 3).
Later I did normalization on this file.
I use fillna to fill the missing data in subject column.
After this I download the Stopwords using nltk package.
Later I prepare the model to remove punctuation, remove stopwords and return the clean text.

Use of Tokenization

I used tokenization to tokenized the spam messages in Data.

spam['message'].head().apply(process_text)

Convert a collection of text to a matrix of tokens

I used count vectorizer to import Bow (Bag of words).

Split the data 80% training and 20% testing

I split the data into train test format to get 80% training and 20% of testing data.

Now shape of Bow is (2876, 64661)

Naïve Bayes for Multinomial NB

I used naïve bayes to get the better Accuracy on the data.

```
The further accuracy on train data I got is 0.9973913043478261
```

Also I used for testing data also.

```
In testing data I got accuracy of test data is 0.9895833333333334
```

Conclusion

Train data and test both these data are better accuracy in spam detection.

So, there is only 2% of data is spam in data folder.