Andy Le - 000805099

Machine Learning – Assignment 3b

Data Description: https://www.kaggle.com/team-ai/spam-text-message-classification?fbclid=IwAR2NvH5gP9I_CbQ3Z1mMJyFWOlPjy4JmYCZq6jj-ywj_qiKJNykx_pV2AbM

Classification:

Class	Number of items
Spam	1494
Not Spam	9650

Split:

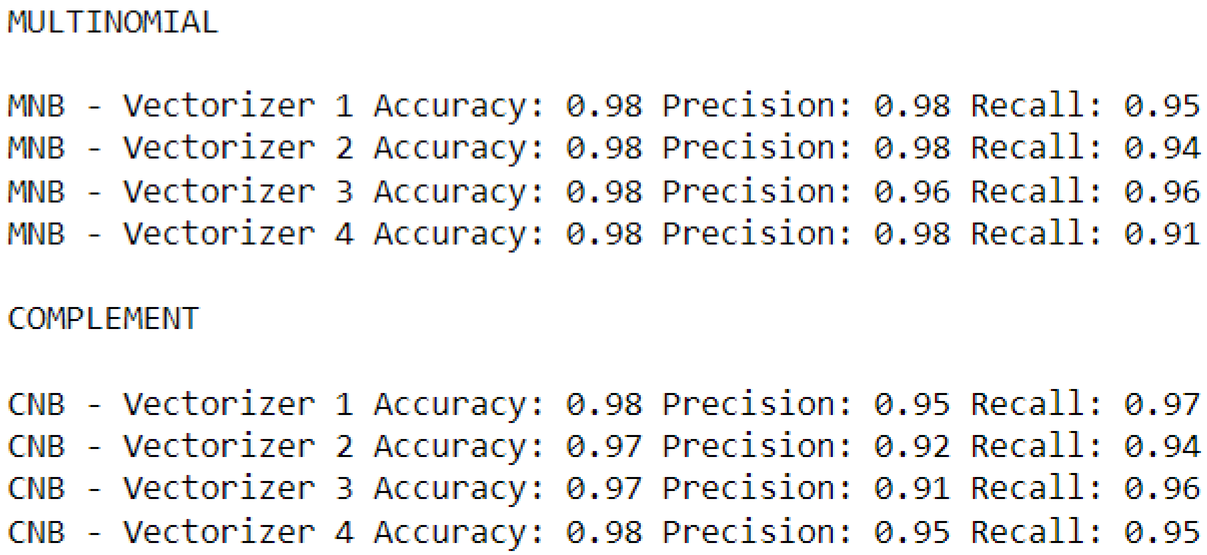
Test size = 30%

Random State = 30

Train set = 3900

Test set = 1672

Results:



```
MULTINOMIAL

MNB - Vectorizer 1 Accuracy: 0.98 Precision: 0.98 Recall: 0.95
MNB - Vectorizer 2 Accuracy: 0.98 Precision: 0.98 Recall: 0.94
MNB - Vectorizer 3 Accuracy: 0.98 Precision: 0.96 Recall: 0.96
MNB - Vectorizer 4 Accuracy: 0.98 Precision: 0.98 Recall: 0.91

COMPLEMENT

CNB - Vectorizer 1 Accuracy: 0.98 Precision: 0.95 Recall: 0.97
CNB - Vectorizer 2 Accuracy: 0.97 Precision: 0.92 Recall: 0.94
CNB - Vectorizer 3 Accuracy: 0.97 Precision: 0.91 Recall: 0.96
CNB - Vectorizer 4 Accuracy: 0.98 Precision: 0.95 Recall: 0.95
```

Discussion:

There are no clear winners overall, though comparing the Complement and Multinomial displays that the text representation is slightly better. This may be since Complement Naïve Bayes is suited for imbalanced data sets, where Multinomial Naïve Bayes is suited for classifications with discrete features. Looking at the bigger picture, the differences in having ngram(2,2) and stop_words(‘English’), is not dramatically different. I would recommend Multinomial with stop-words=(‘English’).

Future Work:

I would like to test if email spam can be comparable to SMS spam.