Programming Language Used: Python

Emails were processed using TF-IDF vectorizer for extracting the features and also finding the significance of each feature in every mail. 1-gram, 2-gram and 3-gram models were used for the feature extraction.

The classification based on these extracted features were done with five different classifiers:

Support Vector Machines

- K-Nearest Neighbors
- Decision Trees
- Random Forest
- Multi-Layer Perceptron

For SVM, 4 different kernels were tried out:

- Linear Kernel
- Radial Basis Kernel
- Polynomial Kernel
- Sigmoid Kernel

For K-Nearest Neighbors, the number of nearest neighbors were varied from 1 to 50 and were found out that the maximum accuracy was around 6-8 neighbors.

For Decision Trees, the maximum depth was treated as a varying parameter.

For Random Forest, the number of base classifiers were varied.

For Multi-Layer Perceptron, the maximum iterations were varied so that the effect of number of queries over the data could be noted.