1. Attempt No 1
    1. Model details:
        - Applied SVM
        - CountVectorizer having 5000 features
        - Training and Test set divided into 80 and 20%
    2. Error Analysis:
        - 2314 examples in cross validation set
        - Algorithm misclassifies 1886 emails
    3. Improvement suggested
        - Stop words can be applied
    4. Precision, Recall, F1-Score:
        - Accuracy: 18.4%
        - Precision: 0.59
        - Recall: 0.502
        - F1-Score: 0.158

2. Attempt No 2
    1. Model details:
        - Applied SVM
        - CountVectorizer having 5000 features
        - Training and Test set divided into 80 and 20%
        - Removed Stop words from corpus
    2. Error Analysis:
        - 2314 examples in cross validation set
        - Algorithm misclassifies 1574 emails
    3. Improvement suggested
        - Stemming can be applied
    4. Precision, Recall, F1-Score:
        - Accuracy: 31.9%
        - Precision: 0.586
        - Recall: 0.573
        - F1-Score: 0.318

3. Attempt No 3
    1. Model details:
        - Applied SVM
        - CountVectorizer having 5000 features
        - Training and Test set divided into 80 and 20%
        - Removed Stop words from corpus
        - Applied PorterStemmer
    2. Error Analysis:
        - 2314 examples in cross validation set
        - Algorithm misclassifies 1874 emails
    3. Improvement suggested
        - Apply LancasterStemmer

   4. Precision, Recall, F1-Score:
      - Accuracy: 18.9%
      - Precision: 0.45
      - Recall: 0.49
      - F1-Score: 0.168

4. Attempt No 4
   1. Model details:
      - Applied SVM
      - CountVectorizer having 5000 features
      - Training and Test set divided into 80 and 20%
      - Removed Stop words from corpus
      - Applied LancasterStemmer
   2. Error Analysis:
      - 2314 examples in cross validation set
      - Algorithm misclassifies 1614 emails
   3. Improvement suggested
      - Only allow non English words
   4. Precision, Recall, F1-Score:
      - Accuracy: 30.2%
      - Precision: 0.46
      - Recall: 0.45
      - F1-Score: 0.3

5. Attempt No 5
   1. Model details:
      - Applied SVM
      - CountVectorizer having 5000 features
      - Training and Test set divided into 80 and 20%
      - Removed Stop words from corpus
      - Removed non English words
   2. Error Analysis:
      - 2314 examples in cross validation set
      - Algorithm misclassifies 1874 emails
   3. Improvement suggested
      - Apply LancasterStemmer and removed non English words
   4. Precision, Recall, F1-Score:
      - Accuracy: 36.6%
      - Precision: 0.47
      - Recall: 0.46
      - F1-Score: 0.35

6. Attempt No 6
   1. Model details:
      - Applied SVM

- CountVectorizer having 5000 features
- Training and Test set divided into 80 and 20%
- Removed Stop words from corpus
- Removed non English words
- Apply Lancaster Stemmer

2. Error Analysis:
   - 2314 examples in cross validation set
   - Algorithm misclassifies 1614 emails
3. Improvement suggested
   - Change model Parameters
4. Precision, Recall, F1-Score:
   - Accuracy: 30.2%
   - Precision: 0.46
   - Recall: 0.45
   - F1-Score: 0.30

7. Attempt No 7
   1. Model details:
      - Applied SVM
      - CountVectorizer having 5000 features
      - Training and Test set divided into 80 and 20%
      - Removed Stop words from corpus
      - Removed non English words
      - Removed gamma=1 parameter with kernel only sigmoid
   2. Error Analysis:
      - 2314 examples in cross validation set
      - Algorithm misclassifies 415 emails
   3. Improvement suggested
      - Further tinker with parameters of model
   4. Precision, Recall, F1-Score:
      - Accuracy: 82.0%
      - Precision: 0.91
      - Recall: 0.50
      - F1-Score: 0.46

8. Attempt No 8
   1. Model details:
      - Applied SVM
      - CountVectorizer having 5000 features
      - Training and Test set divided into 80 and 20%
      - Removed Stop words from corpus
      - Removed non English words
      - Applied poly SVM with degree = 5
   2. Error Analysis:
      - 2314 examples in cross validation set

    - Algorithm misclassifies 419 emails
    3. Improvement suggested
    - Further tinker with parameters of model
    4. Precision, Recall, F1-Score:
    - Accuracy: 81.0%
    - Precision: 0.40
    - Recall: 0.5
    - F1-Score: 0.45

9. Attempt No 9
    1. Model details:
    - Applied SVM
    - Can't hard code features as it will not work with test prediction
    - Works with very low features in vectorisation such as 50 but precision is low now
    - Used Tfidf to solve this issue
    - Removed non English words
    - Applied svm with sigmoid kernel
    2. Error Analysis:
    - 2314 examples in cross validation set
    - Algorithm misclassifies 419 emails
    3. Improvement suggested
    - Further tinker with parameters of model
    4. Precision, Recall, F1-Score:
    - Accuracy: 81.0%
    - Precision: 0.40
    - Recall: 0.5
    - F1-Score: 0.45

10. Attempt No 10
    1. Model details:
    - I found out that my all my previous values for precision, recall, f1score, miss-classified emails are incorrect coding wise
    - I used parameter with CountVectorizer with min_df =0.0001 and max_df =0.8 these parameters were decided by making intelligent decision making.
    - I found out that when making prediction only transform is applied
    - Removed single character letters
    - Mostly pre processing is done when max_df is 0.8 which mean ignore terms that are present in 80% of the document
    2. Error Analysis:
    - 579 examples in cross validation set
    - Algorithm misclassifies 50 emails
    3. Improvement suggested

- Now algorithm can be further improved by trying out different kernels of svm

4. Precision, Recall, F1-Score:

- Accuracy: 91.1%
- Precision: 0.974
- Recall: 0.921
- F1-Score: 0.947
- Confusion matrix:
  [462 12]
  [39 66]