Appendix Contents

Xu Jun Zhu Bellmore data – no pre-processing

```
===============================================================
KNN classifier results
-------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.93      0.85       162
        Yes       0.92      0.77      0.84       186

avg / total       0.86      0.84      0.84       348

The accuracy score is 84.48%
--------------------------
Confusion Matrix
--------------------------
[[150  12]
 [ 42 144]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.916981946104
fpr:  [ 0.          0.00617284  0.07407407  0.17901235  0.5         1.        ]
tpr:  [ 0.53225806  0.62903226  0.77419355  0.88709677  0.94086022  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
===============================================================
SVC classifier results
-------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.97      0.89       162
        Yes       0.97      0.83      0.89       186

avg / total       0.90      0.89      0.89       348

The accuracy score is 89.37%
--------------------------
Confusion Matrix
--------------------------
[[157   5]
 [ 32 154]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.965717509624
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.02688172 ...,  1.          1.          1.        ]
threshold:  [ 0.99999837  0.99999719  0.99998746 ...,  0.00505684  0.00407533
  0.00172068]
===============================================================
Decision Tree classifier results
-------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.81      0.99      0.89       162
        Yes       0.99      0.80      0.89       186

avg / total       0.91      0.89      0.89       348

The accuracy score is 89.08%
--------------------------
```

```
Confusion Matrix
---------------------------
[[161   1]
 [ 37 149]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.893767423337
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.80107527  0.80107527  0.80107527  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.98      0.90       162
        Yes       0.98      0.82      0.89       186

avg / total       0.91      0.90      0.90       348

The accuracy score is 89.66%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.971790787203
fpr:  [ 0.          0.          0.         ...,  0.45679012  0.47530864  1.        ]
tpr:  [ 0.43010753  0.44086022  0.55376344 ...,  0.97849462  0.97849462  1.        ]
threshold:  [ 1.          0.95        0.9        ...,  0.05        0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.83      0.84       162
        Yes       0.86      0.87      0.86       186

avg / total       0.85      0.85      0.85       348

The accuracy score is 85.06%
---------------------------
Confusion Matrix
---------------------------
[[135  27]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.942951015532
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
tpr:  [ 0.24193548  0.25806452  0.29569892 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   1.20722861e-27
   2.26637291e-29   8.14530906e-36]
=================================================================
Optimized SVC hyper parameters
```

```
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.98      0.94       162
        Yes       0.98      0.90      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.68%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 19 167]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980651798752
fpr:  [ 0.          0.          0.         ...,  0.9382716   0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03225806 ...,  1.          1.          1.        ]
threshold:  [ 0.99999757  0.99676015  0.994483   ...,  0.04764662  0.03219192
  0.02867912]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       162
        Yes       0.97      0.91      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 16 170]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.986509358821
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.96296296  1.        ]
tpr:  [ 0.20430108  0.2688172   0.27419355 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.98888889  0.98333333 ...,  0.01111111  0.00555556  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.90       162
        Yes       0.99      0.82      0.90       186
```

```
avg / total       0.92      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.912883313421
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.8172043   0.82258065  0.83333333  0.83870968  0.83870968
  1.        ]
threshold:  [ 2.          1.          0.66666667  0.5         0.33333333  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.96      0.91       162
        Yes       0.96      0.86      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.80%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 26 160]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.911589008363
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.86021505  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 11.264480829238892 seconds
================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with porter stemmer

```
================================================================
KNN classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.93      0.85       162
        Yes       0.92      0.77      0.84       186

avg / total       0.86      0.84      0.84       348
```

```
The accuracy score is 84.48%
---------------------------
Confusion Matrix
---------------------------
[[150  12]
 [ 42 144]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.916981946104
fpr:  [ 0.          0.00617284  0.07407407  0.17901235  0.5         1.        ]
tpr:  [ 0.53225806  0.62903226  0.77419355  0.88709677  0.94086022  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.97      0.89       162
        Yes       0.97      0.83      0.89       186

avg / total       0.90      0.89      0.89       348

The accuracy score is 89.37%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 32 154]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965717509624
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.02688172 ...,  1.          1.          1.        ]
threshold:  [ 0.99999837  0.99999719  0.99998746 ...,  0.00505684  0.00407533
  0.00172068]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.99      0.93       162
        Yes       0.99      0.88      0.93       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 93.39%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 22 164]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.938188636665
fpr:  [ 0.          0.00617284  0.03703704  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.88172043  0.88172043  0.88709677  0.88709677  1.        ]
```

```
threshold:  [ 2.          1.          0.5         0.33333333  0.25        0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.98      0.90       162
        Yes       0.98      0.82      0.89       186

avg / total       0.91      0.90      0.90       348

The accuracy score is 89.66%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.971790787203
fpr:  [ 0.          0.          0.         ...,  0.45679012  0.47530864  1.        ]
tpr:  [ 0.43010753  0.44086022  0.55376344 ...,  0.97849462  0.97849462  1.        ]
threshold:  [ 1.          0.95        0.9        ...,  0.05        0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.83      0.84       162
        Yes       0.86      0.87      0.86       186

avg / total       0.85      0.85      0.85       348

The accuracy score is 85.06%
---------------------------
Confusion Matrix
---------------------------
[[135  27]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.942951015532
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
tpr:  [ 0.24193548  0.25806452  0.29569892 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   1.20722861e-27
   2.26637291e-29   8.14530906e-36]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.98      0.94       162
        Yes       0.98      0.90      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.68%
---------------------------
```

```
Confusion Matrix
---------------------------
[[159   3]
 [ 19 167]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980651798752
fpr:  [ 0.          0.          0.         ...,  0.9382716   0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03225806 ...,  1.          1.          1.        ]
threshold:  [ 0.99999757  0.99676015  0.994483   ...,  0.04764662  0.03219192
  0.02867912]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       162
        Yes       0.97      0.91      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 16 170]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.986509358821
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.96296296  1.        ]
tpr:  [ 0.20430108  0.2688172   0.27419355 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.98888889  0.98333333 ...,  0.01111111  0.00555556  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.91       162
        Yes       0.99      0.84      0.91       186

avg / total       0.93      0.91      0.91       348

The accuracy score is 91.38%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 29 157]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.918657905217
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.84408602  0.84408602  0.84946237  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.96      0.91       162
        Yes       0.96      0.86      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.80%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 26 160]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.911589008363
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.86021505  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 11.264480829238892 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with lemmatizer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.77      0.91      0.84       162
        Yes       0.91      0.77      0.83       186

avg / total       0.84      0.83      0.83       348

The accuracy score is 83.33%
---------------------------
Confusion Matrix
---------------------------
[[147  15]
 [ 43 143]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.909780300013
```

```
fpr:  [ 0.          0.01234568  0.09259259  0.16049383  0.48148148  1.        ]
tpr:  [ 0.50537634  0.59139785  0.7688172   0.87634409  0.93548387  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.91       162
        Yes       0.99      0.85      0.91       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 91.38%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 28 158]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.968339307049
fpr:  [ 0.          0.          0.         ...,  0.9691358   0.99382716  1.        ]
tpr:  [ 0.00537634  0.02688172  0.04301075 ...,  1.          1.          1.        ]
threshold:  [  9.99997174e-01   9.99995830e-01   9.99986288e-01 ...,   4.00891683e-03
   2.71443542e-03   6.52049230e-04]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.91       162
        Yes       0.99      0.83      0.90       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 90.52%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 32 154]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.91031129696
fpr:  [ 0.          0.00617284  0.03703704  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.82795699  0.82795699  0.83333333  0.83333333  1.        ]
threshold:  [ 2.          1.          0.5         0.33333333  0.25        0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.97      0.90       162
        Yes       0.97      0.83      0.90       186

avg / total       0.91      0.90      0.90       348
```

```
The accuracy score is 89.66%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 31 155]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.9741304925
fpr:  [ 0.          0.          0.00617284 ...,  0.52469136  0.53703704  1.        ]
tpr:  [ 0.44623656  0.4516129   0.55376344 ...,  0.98387097  0.98387097  1.        ]
threshold:  [ 1.      0.95    0.9    ...,  0.05    0.0375  0.    ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.84      0.84       162
        Yes       0.86      0.86      0.86       186

avg / total       0.85      0.85      0.85       348

The accuracy score is 85.06%
---------------------------
Confusion Matrix
---------------------------
[[136  26]
 [ 26 160]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.941723085092
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
tpr:  [ 0.23655914  0.24193548  0.25268817 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   6.01752732e-27
   6.32188224e-28   1.10837477e-34]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.99      0.94       162
        Yes       0.99      0.91      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.985497145891
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.00537634  0.02688172  0.04301075 ...,  1.          1.          1.        ]
threshold:  [ 0.99999876  0.99681433  0.99548642 ...,  0.02995601  0.02010554
```

```
  0.01913084]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None, 'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
============================================================
Optimized Random Forest hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.99      0.94       162
        Yes       0.99      0.90      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 18 168]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980668392407
fpr:  [ 0.          0.          0.         ...,  0.91358025  0.95061728  1.        ]
tpr:  [ 0.21505376  0.29032258  0.33333333 ...,  0.99462366  0.99462366  1.        ]
threshold:  [ 1.          0.98888889  0.97777778 ...,  0.0125      0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
============================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.99      0.91       162
        Yes       0.99      0.84      0.91       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 91.09%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 30 156]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.923702376211
fpr:  [ 0.          0.00617284  0.03703704  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.83870968  0.83870968  0.86021505  0.86021505  1.        ]
threshold:  [ 2.          1.          0.5         0.33333333  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
============================================================
Optimized KNN hyper parameters
------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.84      0.96      0.90       162
        Yes       0.96      0.84      0.90       186

avg / total       0.91      0.90      0.90       348

The accuracy score is 89.66%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 30 156]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.900836320191
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.83870968  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 15.492542028427124 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with tri-grams

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.91      0.84       162
        Yes       0.91      0.78      0.84       186

avg / total       0.85      0.84      0.84       348

The accuracy score is 83.91%
---------------------------
Confusion Matrix
---------------------------
[[147  15]
 [ 41 145]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.908817868047
fpr:  [ 0.          0.01851852  0.09259259  0.19753086  0.46296296  1.        ]
tpr:  [ 0.53225806  0.62903226  0.77956989  0.88172043  0.93010753  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.99      0.92       162
        Yes       0.99      0.85      0.92       186
```

```
avg / total       0.93      0.92      0.92       348

The accuracy score is 91.95%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965783884243
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.02150538  0.03225806  0.07526882 ...,  1.          1.          1.        ]
threshold:  [  9.99998258e-01   9.99998111e-01   9.99985815e-01 ...,   3.06500934e-03
   2.79019800e-03   7.62335019e-04]
====================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.90       162
        Yes       0.99      0.82      0.90       186

avg / total       0.92      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.904918359219
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.82258065  0.82258065  0.82258065  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
====================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.98      0.90       162
        Yes       0.97      0.82      0.89       186

avg / total       0.91      0.89      0.89       348

The accuracy score is 89.37%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.974827425992
```

```
fpr:  [ 0.          0.00617284  0.00617284 ...,  0.4382716   0.45061728  1.        ]
tpr:  [ 0.          0.40322581  0.41935484 ...,  0.98387097  0.98387097  1.        ]
threshold:  [ 2.          1.          0.91111111 ...,  0.06666667  0.03333333  0.        ]
==================================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.84      0.84       162
        Yes       0.86      0.87      0.86       186

avg / total       0.85      0.85      0.85       348

The accuracy score is 85.34%
---------------------------
Confusion Matrix
---------------------------
[[136  26]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943747510952
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
tpr:  [ 0.24193548  0.24731183  0.25806452 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   7.99772054e-28
   2.41964927e-28   7.23788335e-35]
==================================================================
Optimized SVC hyper parameters
--------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.989114562591
fpr:  [ 0.          0.          0.         ...,  0.90123457  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03763441 ...,  1.          1.          1.        ]
threshold:  [ 0.99999739  0.99650344  0.99613449 ...,  0.04769611  0.02495623
  0.02426545]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
--------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.99      0.94       162
        Yes       0.99      0.91      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980286738351
fpr:  [ 0.          0.          0.         ...,  0.89506173  0.94444444  1.        ]
tpr:  [ 0.16666667  0.17741935  0.19892473 ...,  0.99462366  1.          1.        ]
threshold:  [ 1.          0.99444444  0.99166667 ...,  0.01555556  0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.92       162
        Yes       0.99      0.85      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.67%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 28 158]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.918857029072
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.84946237  0.84946237  0.84946237  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.96      0.90       162
        Yes       0.96      0.85      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.52%
---------------------------
Confusion Matrix
```

```
---------------------------
[[156   6]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.90890083632
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.85483871  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 12.598369121551514 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with bi-grams

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.91      0.84       162
        Yes       0.91      0.78      0.84       186

avg / total       0.85      0.84      0.84       348

The accuracy score is 83.91%
---------------------------
Confusion Matrix
---------------------------
[[147  15]
 [ 41 145]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.908817868047
fpr:  [ 0.          0.01851852  0.09259259  0.19753086  0.46296296  1.        ]
tpr:  [ 0.53225806  0.62903226  0.77956989  0.88172043  0.93010753  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.99      0.92       162
        Yes       0.99      0.85      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.95%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.965783884243
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.02150538  0.03225806  0.07526882 ...,  1.          1.          1.        ]
threshold:  [  9.99997545e-01   9.99997342e-01   9.99980982e-01 ...,   3.52612044e-03
   3.21723203e-03   9.06874568e-04]
================================================================
Decision Tree classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.90       162
        Yes       0.99      0.82      0.90       186

avg / total       0.92      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.904918359219
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.82258065  0.82258065  0.82258065  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
================================================================
Random Forest classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.98      0.88       162
        Yes       0.98      0.79      0.87       186

avg / total       0.90      0.88      0.88       348

The accuracy score is 87.93%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 39 147]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978295499801
fpr:  [ 0.          0.00617284  0.00617284 ...,  0.19753086  0.45061728  1.        ]
tpr:  [ 0.38709677  0.58064516  0.68817204 ...,  0.98387097  0.98924731  1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
================================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.84      0.84       162
        Yes       0.86      0.87      0.86       186

avg / total       0.85      0.85      0.85       348
```

```
The accuracy score is 85.34%
---------------------------
Confusion Matrix
---------------------------
[[136  26]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943747510952
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
tpr:  [ 0.24193548  0.24731183  0.25806452 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   7.99772054e-28
   2.41964927e-28   7.23788335e-35]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.989114562591
fpr:  [ 0.          0.          0.         ...,  0.90123457  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03763441 ...,  1.          1.          1.        ]
threshold:  [ 0.99999949  0.9999906   0.99998827 ...,  0.03950223  0.01930683
  0.01871801]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 17 169]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.980170582769
fpr:  [ 0.          0.          0.         ...,  0.92592593  0.9382716   1.        ]
tpr:  [ 0.1344086   0.13978495  0.28494624 ...,  0.99462366  0.99462366  1.        ]
threshold:  [ 1.          0.99444444  0.98888889 ...,  0.01111111  0.00277778  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
====================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.92       162
        Yes       0.99      0.85      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.67%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 28 158]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.918857029072
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.84946237  0.84946237  0.84946237  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
====================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.96      0.90       162
        Yes       0.96      0.85      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.52%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.90890083632
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.85483871  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}
```

```
=================================================================
Program ran in 12.437262535095215 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with 4-grams

```
=================================================================
KNN classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.91      0.84       162
        Yes       0.91      0.78      0.84       186

avg / total       0.85      0.84      0.84       348

The accuracy score is 83.91%
---------------------------
Confusion Matrix
---------------------------
[[147  15]
 [ 41 145]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.908817868047
fpr:  [ 0.          0.01851852  0.09259259  0.19753086  0.46296296  1.        ]
tpr:  [ 0.53225806  0.62903226  0.77956989  0.88172043  0.93010753  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.99      0.92       162
        Yes       0.99      0.85      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.95%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965783884243
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.02150538  0.03225806  0.07526882 ...,  1.          1.          1.        ]
threshold:  [  9.99998341e-01   9.99998200e-01   9.99986495e-01 ...,   3.14915787e-03
   2.86686046e-03   7.83499215e-04]
=================================================================
Decision Tree classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.90       162
        Yes       0.99      0.82      0.90       186
```

```
avg / total       0.92      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.904918359219
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.82258065  0.82258065  0.82258065  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.98      0.90       162
        Yes       0.97      0.82      0.89       186

avg / total       0.91      0.89      0.89       348

The accuracy score is 89.37%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.975590734103
fpr:  [ 0.          0.00617284  0.00617284 ...,  0.43209877  0.44444444  1.        ]
tpr:  [ 0.          0.48387097  0.48924731 ...,  0.98924731  0.98924731  1.        ]
threshold:  [ 2.          1.          0.96666667 ...,  0.05        0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.84      0.84       162
        Yes       0.86      0.87      0.86       186

avg / total       0.85      0.85      0.85       348

The accuracy score is 85.34%
---------------------------
Confusion Matrix
---------------------------
[[136  26]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943747510952
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.97530864  1.        ]
```

```
tpr:  [ 0.24193548  0.24731183  0.25806452 ...,  0.99462366  1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   7.99772054e-28
   2.41964927e-28   7.23788335e-35]
=================================================================
Optimized SVC hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.989114562591
fpr:  [ 0.          0.          0.         ...,  0.90123457  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03763441 ...,  1.          1.          1.        ]
threshold:  [ 0.99999732  0.99645186  0.99607716 ...,  0.04668964  0.02440942
  0.02373308]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.98      0.93       162
        Yes       0.98      0.89      0.93       186

avg / total       0.93      0.93      0.93       348

The accuracy score is 92.82%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 21 165]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980767954334
fpr:  [ 0.          0.          0.         ...,  0.88888889  0.93209877  1.        ]
tpr:  [ 0.18817204  0.23655914  0.32795699 ...,  0.99462366  1.          1.        ]
threshold:  [ 1.          0.98888889  0.97777778 ...,  0.02222222  0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.85      0.99      0.92       162
        Yes       0.99      0.85      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.67%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 28 158]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.918857029072
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.84946237  0.84946237  0.84946237  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.96      0.90       162
        Yes       0.96      0.85      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.52%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.90890083632
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.85483871  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1, 'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 14.780819416046143 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with stop word removal

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      0.49      0.63       162
```

```
        Yes       0.68      0.94      0.79       186

avg / total       0.77      0.73      0.71       348

The accuracy score is 72.99%
---------------------------
Confusion Matrix
---------------------------
[[ 80  82]
 [ 12 174]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.929974777645
fpr:  [ 0.          0.16049383  0.50617284  0.64197531  0.79012346  1.        ]
tpr:  [ 0.84946237  0.87634409  0.93548387  0.96236559  0.96774194  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.88      0.86       162
        Yes       0.89      0.85      0.87       186

avg / total       0.87      0.87      0.87       348

The accuracy score is 86.78%
---------------------------
Confusion Matrix
---------------------------
[[143  19]
 [ 27 159]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950335191823
fpr:  [ 0.          0.          0.         ...,  0.98765432  0.98765432  1.        ]
tpr:  [ 0.00537634  0.01075269  0.02150538 ...,  0.99462366  1.          1.        ]
threshold:  [ 0.99176514  0.99167323  0.99153655 ...,  0.0230722   0.02249433
  0.01108441]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.81      0.99      0.89       162
        Yes       0.99      0.80      0.89       186

avg / total       0.91      0.89      0.89       348

The accuracy score is 89.08%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 37 149]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.893767423337
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.80107527  0.80107527  0.80107527  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
==============================================================
Random Forest classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.98      0.93       162
        Yes       0.98      0.89      0.93       186

avg / total       0.93      0.93      0.93       348

The accuracy score is 92.82%
--------------------------
Confusion Matrix
--------------------------
[[158   4]
 [ 21 165]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.97149210142
fpr:  [ 0.          0.          0.00617284 ...,  0.42592593  0.44444444  1.        ]
tpr:  [ 0.67741935  0.68817204  0.75268817 ...,  0.98924731  0.98924731  1.        ]
threshold:  [ 1.          0.93333333  0.9        ...,  0.1         0.025       0.        ]
==============================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.78      0.86       162
        Yes       0.84      0.97      0.90       186

avg / total       0.89      0.88      0.88       348

The accuracy score is 88.22%
--------------------------
Confusion Matrix
--------------------------
[[127  35]
 [  6 180]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.970662418691
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.98148148  1.        ]
tpr:  [ 0.01075269  0.04301075  0.06451613 ...,  0.99462366  1.          1.        ]
threshold:  [  9.99999721e-01   9.99999133e-01   9.99997889e-01 ...,   1.33553927e-07
   9.55331234e-08   5.36638168e-09]
==============================================================
Optimized SVC hyper parameters
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.94      0.94      0.94       348
```

```
The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978262312492
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.01075269  0.02150538  0.03763441 ...,  1.          1.          1.        ]
threshold:  [ 0.99286685  0.99251195  0.9908462  ...,  0.05372672  0.0537254
  0.05372081]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
======================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.97      0.95       162
        Yes       0.97      0.94      0.96       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 95.40%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 11 175]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.979224744458
fpr:  [ 0.          0.          0.         ...,  0.94444444  0.96296296  1.        ]
tpr:  [ 0.2688172   0.33333333  0.33870968 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.98888889  0.98055556 ...,  0.01269841  0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
======================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.74      0.99      0.85       162
        Yes       0.99      0.70      0.82       186

avg / total       0.88      0.84      0.83       348

The accuracy score is 83.62%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
```

```
 [ 56 130]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.873058542413
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.69892473  0.76344086  0.76344086  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.96      0.94       162
        Yes       0.97      0.93      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 13 173]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.946535244922
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.93010753  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 9.370222568511963 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with TF-IDF

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.77      0.81      0.79       162
        Yes       0.83      0.79      0.81       186

avg / total       0.80      0.80      0.80       348

The accuracy score is 79.89%
---------------------------
Confusion Matrix
---------------------------
[[131  31]
 [ 39 147]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.903308774725
fpr:  [ 0.          0.0617284   0.19135802  0.27160494  0.63580247  1.        ]
tpr:  [ 0.61827957  0.69354839  0.79032258  0.91935484  0.95698925  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.98      0.91       162
        Yes       0.98      0.84      0.90       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 90.52%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 30 156]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973051904952
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.98765432  1.        ]
tpr:  [ 0.00537634  0.03225806  0.04301075 ...,  1.          1.          1.        ]
threshold:  [ 0.99999988  0.9957604   0.99445964 ...,  0.01890579  0.01496237
  0.00407099]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.99      0.90       162
        Yes       0.99      0.82      0.90       186

avg / total       0.92      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 33 153]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.904918359219
fpr:  [ 0.          0.00617284  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.82258065  0.82258065  0.82258065  1.        ]
threshold:  [ 2.    1.    0.5   0.25  0.  ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.99      0.91       162
        Yes       0.99      0.84      0.91       186
```

```
avg / total       0.92      0.91      0.91       348

The accuracy score is 91.09%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 30 156]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.971790787203
fpr:  [ 0.          0.          0.         ...,  0.40740741  0.40740741  1.        ]
tpr:  [ 0.48924731  0.5         0.51075269 ...,  0.97311828  0.97849462  1.        ]
threshold:  [ 1.          0.98333333  0.95714286 ...,  0.1         0.05        0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.79      0.99      0.88       162
        Yes       0.99      0.77      0.87       186

avg / total       0.90      0.87      0.87       348

The accuracy score is 87.36%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 42 144]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.964821452277
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.97530864  1.        ]
tpr:  [ 0.2311828   0.23655914  0.26344086 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   8.84825245e-24
   7.08926980e-25   7.58123885e-37]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 16 170]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.979257931767
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
```

```
tpr:  [ 0.00537634  0.06989247  0.08064516 ...,  1.          1.          1.        ]
threshold:  [ 0.99998711  0.98907005  0.98906865 ...,  0.03809055  0.03808744
  0.03808424]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
===============================================================
Optimized Random Forest hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       162
        Yes       0.98      0.92      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 95.11%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 14 172]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.981199389354
fpr:  [ 0.          0.          0.         ...,  0.85185185  0.93209877  1.        ]
tpr:  [ 0.18817204  0.19354839  0.20430108 ...,  0.99462366  1.          1.        ]
threshold:  [ 1.          0.99777778  0.99166667 ...,  0.01296296  0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
===============================================================
Optimized Decision Tree hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.99      0.94       162
        Yes       0.99      0.90      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 18 168]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.951928182663
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.89784946  0.90322581  0.90322581  0.91397849  0.91397849
  1.        ]
threshold:  [ 2.          1.          0.66666667  0.5         0.33333333  0.25        0.        ]
```

```
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
===================================================================
Optimized KNN hyper parameters
-------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.96      0.90       162
        Yes       0.96      0.85      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.52%
----------------------------
Confusion Matrix
----------------------------
[[156   6]
 [ 27 159]]
----------------------------
ROC Curve
----------------------------
Area under the curve:  0.90890083632
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.85483871  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

===================================================================
Program ran in 8.42959976196289 seconds
===================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with TF-IDF + Tri-grams

```
===================================================================
KNN classifier results
-------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.81      0.78      0.80       162
        Yes       0.82      0.84      0.83       186

avg / total       0.82      0.82      0.82       348

The accuracy score is 81.61%
----------------------------
Confusion Matrix
----------------------------
[[127  35]
 [ 29 157]]
----------------------------
ROC Curve
----------------------------
Area under the curve:  0.90249568565
fpr:  [ 0.          0.07407407  0.21604938  0.34567901  0.64197531  1.        ]
tpr:  [ 0.61290323  0.6827957   0.84408602  0.91935484  0.96774194  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
===================================================================
SVC classifier results
```

```
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      0.97      0.90       162
        Yes       0.97      0.84      0.90       186

avg / total       0.91      0.90      0.90       348

The accuracy score is 90.23%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 29 157]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.969965485198
fpr:  [ 0.          0.          0.         ...,  0.9691358   0.99382716  1.        ]
tpr:  [ 0.00537634  0.05376344  0.06451613 ...,  1.          1.          1.        ]
threshold:  [ 0.99999985  0.99591566  0.99581357 ...,  0.01343539  0.01065648
  0.00860201]
=================================================================
Decision Tree classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      0.99      0.93       162
        Yes       0.99      0.87      0.93       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 92.82%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 24 162]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.940495154653
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.8655914   0.87096774  0.87634409  0.89247312  0.89247312
  1.        ]
threshold:  [ 2.          1.          0.66666667  0.5         0.33333333  0.25        0.        ]
=================================================================
Random Forest classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.99      0.92       162
        Yes       0.99      0.86      0.92       186

avg / total       0.93      0.92      0.92       348

The accuracy score is 91.95%
---------------------------
Confusion Matrix
```

```
---------------------------
[[160   2]
 [ 26 160]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.968538430904
fpr:  [ 0.          0.          0.         ...,  0.39506173  0.40740741  1.        ]
tpr:  [ 0.48387097  0.49462366  0.67741935 ...,  0.97311828  0.97311828  1.        ]
threshold:  [ 1.          0.93333333  0.9        ...,  0.06        0.02857143  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.81      0.99      0.89       162
        Yes       0.99      0.80      0.88       186

avg / total       0.90      0.89      0.88       348

The accuracy score is 88.51%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 38 148]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.962631089871
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.97530864  1.        ]
tpr:  [ 0.30107527  0.31182796  0.33333333 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   4.58613458e-37
   2.49283914e-38   1.87270870e-55]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.97      0.93       162
        Yes       0.97      0.91      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.68%
---------------------------
Confusion Matrix
---------------------------
[[157   5]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.980917297226
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.95679012  1.        ]
tpr:  [ 0.00537634  0.0483871   0.05913978 ...,  0.99462366  1.          1.        ]
threshold:  [ 0.99705935  0.9869792   0.98671013 ...,  0.04718283  0.04718143
  0.04716843]
```

```
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.94       162
        Yes       0.98      0.91      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 94.25%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 17 169]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978992433294
fpr:  [ 0.          0.          0.         ...,  0.94444444  0.94444444  1.        ]
tpr:  [ 0.24193548  0.25268817  0.33333333 ...,  0.99462366  1.          1.        ]
threshold:  [ 1.          0.99259259  0.98888889 ...,  0.01111111  0.00444444  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.99      0.93       162
        Yes       0.99      0.88      0.93       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 93.39%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 22 164]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943382450551
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.87634409  0.88172043  0.88172043  0.89784946  0.89784946
  1.        ]
threshold:  [ 2.          1.          0.66666667  0.5         0.33333333  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
```

```
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      0.96      0.91       162
        Yes       0.96      0.88      0.92       186

avg / total       0.92      0.92      0.92       348

The accuracy score is 91.67%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 23 163]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.919653524492
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.87634409  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 11.456607103347778 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with TF-IDF + porter stemmer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.77      0.90      0.83       162
        Yes       0.90      0.77      0.83       186

avg / total       0.84      0.83      0.83       348

The accuracy score is 83.05%
---------------------------
Confusion Matrix
---------------------------
[[146  16]
 [ 43 143]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.90369042878
fpr:  [ 0.          0.0308642   0.09876543  0.24691358  0.63580247  1.        ]
tpr:  [ 0.6344086   0.69354839  0.7688172   0.88709677  0.94623656  1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.91       162
        Yes       0.99      0.84      0.91       186
```

```
avg / total       0.92      0.91      0.91       348

The accuracy score is 91.09%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 29 157]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.971956723749
fpr:  [ 0.          0.          0.         ...,  0.97530864  0.98765432  1.        ]
tpr:  [ 0.00537634  0.05913978  0.06989247 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.99738511  0.99643807 ...,  0.0078344   0.00658339
  0.00440047]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.99      0.94       162
        Yes       0.99      0.89      0.94       186

avg / total       0.95      0.94      0.94       348

The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 20 166]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943764104606
fpr:  [ 0.          0.00617284  0.03703704  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.89247312  0.89247312  0.89784946  0.89784946  1.        ]
threshold:  [ 2.          1.          0.5         0.33333333  0.25        0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.99      0.91       162
        Yes       0.99      0.85      0.91       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 91.38%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 28 158]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.972919155715
fpr:  [ 0.          0.00617284  0.00617284 ...,  0.46296296  0.4691358   1.        ]
```

```
tpr:  [ 0.          0.52688172  0.53225806 ...,  0.97849462  0.97849462  1.        ]
threshold:  [ 2.          1.          0.96666667 ...,  0.1         0.05        0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.99      0.88       162
        Yes       0.99      0.78      0.87       186

avg / total       0.90      0.88      0.88       348

The accuracy score is 87.64%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 41 145]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.964456391876
fpr:  [ 0.          0.          0.         ...,  0.96296296  0.97530864  1.        ]
tpr:  [ 0.25268817  0.25806452  0.2688172  ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   1.76125244e-24
   7.16162054e-25   7.03693537e-36]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.98      0.93       162
        Yes       0.98      0.90      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.68%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 18 168]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.982277976902
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.99382716  1.        ]
tpr:  [ 0.00537634  0.08602151  0.09677419 ...,  1.          1.          1.        ]
threshold:  [ 0.99999969  0.99061281  0.99061246 ...,  0.03320055  0.033199    0.03319271]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.94       162
        Yes       0.98      0.91      0.95       186
```

```
avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 16 170]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.981066640117
fpr:  [ 0.          0.          0.         ...,  0.91358025  0.96296296  1.        ]
tpr:  [ 0.19354839  0.20430108  0.29032258 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.99444444  0.98888889 ...,  0.01388889  0.01111111  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.99      0.94       162
        Yes       0.99      0.90      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 94.54%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 18 168]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.951944776318
fpr:  [ 0.          0.00617284  0.03703704  0.03703704  0.04320988  1.        ]
tpr:  [ 0.          0.90322581  0.90322581  0.91397849  0.91397849  1.        ]
threshold:  [ 2.          1.          0.5         0.33333333  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2',
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': 107, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      0.96      0.91       162
        Yes       0.96      0.88      0.92       186

avg / total       0.92      0.92      0.92       348

The accuracy score is 91.67%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
```

```
 [ 23 163]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.919653524492
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.87634409  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 10.741135358810425 seconds
=================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with TF-IDF + lemmatizer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.76      0.84      0.80       162
        Yes       0.85      0.77      0.81       186

avg / total       0.81      0.80      0.80       348

The accuracy score is 80.46%
---------------------------
Confusion Matrix
---------------------------
[[136  26]
 [ 42 144]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.899508827824
fpr:  [ 0.          0.04320988  0.16049383  0.29012346  0.64814815  1.        ]
tpr:  [ 0.62365591  0.68817204  0.77419355  0.90860215  0.9516129   1.        ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.85      0.98      0.91       162
        Yes       0.98      0.84      0.91       186

avg / total       0.92      0.91      0.91       348

The accuracy score is 90.80%
---------------------------
Confusion Matrix
---------------------------
[[159   3]
 [ 29 157]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965817071552
fpr:  [ 0.          0.          0.         ...,  0.98148148  0.98148148  1.        ]
```

```
tpr:  [ 0.00537634  0.04301075  0.05376344 ...,  0.99462366  1.          1.        ]
threshold:  [ 0.99999859  0.99492809  0.99481871 ...,  0.01687658  0.01548019
  0.00348042]
============================================================
Decision Tree classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.99      0.93       162
        Yes       0.99      0.88      0.93       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 93.10%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 23 163]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.93538430904
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.87096774  0.87634409  0.87634409  0.88172043  0.88172043
  1.        ]
threshold:  [ 2.          1.          0.625       0.5         0.33333333  0.25        0.        ]
============================================================
Random Forest classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.99      0.94       162
        Yes       0.99      0.90      0.94       186

avg / total       0.94      0.94      0.94       348

The accuracy score is 93.97%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 19 167]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.975441391212
fpr:  [ 0.          0.00617284  0.00617284 ...,  0.45061728  0.46296296  1.        ]
tpr:  [ 0.          0.43548387  0.44623656 ...,  0.98387097  0.98387097  1.        ]
threshold:  [ 2.          1.          0.96       ...,  0.03333333  0.025       0.        ]
============================================================
Bernoulli Naive Bayes classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.79      0.99      0.88       162
        Yes       0.99      0.77      0.86       186

avg / total       0.89      0.87      0.87       348
```

```
The accuracy score is 87.07%
---------------------------
Confusion Matrix
---------------------------
[[160   2]
 [ 43 143]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.964224080712
fpr:  [ 0.          0.          0.         ...,  0.95679012  0.97530864  1.        ]
tpr:  [ 0.2311828   0.23655914  0.26344086 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   4.98418177e-24
   5.88875062e-25   2.18090566e-34]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.98      0.93       162
        Yes       0.98      0.90      0.94       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 93.39%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 19 167]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.988450816408
fpr:  [ 0.          0.          0.         ...,  0.93209877  0.94444444  1.        ]
tpr:  [ 0.00537634  0.06989247  0.08064516 ...,  1.          1.          1.        ]
threshold:  [ 0.99674204  0.98620247  0.98620043 ...,  0.04548014  0.04547849
  0.04168137]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       162
        Yes       0.98      0.93      0.95       186

avg / total       0.95      0.95      0.95       348

The accuracy score is 95.11%
---------------------------
Confusion Matrix
---------------------------
[[158   4]
 [ 13 173]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.985546926855
fpr:  [ 0.          0.          0.         ...,  0.93209877  0.9382716   1.        ]
tpr:  [ 0.16666667  0.17741935  0.18817204 ...,  1.          1.          1.        ]
threshold:  [ 1.          0.99555556  0.9892284  ...,  0.01111111  0.00555556  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
==================================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      0.99      0.93       162
        Yes       0.99      0.87      0.93       186

avg / total       0.94      0.93      0.93       348

The accuracy score is 92.82%
---------------------------
Confusion Matrix
---------------------------
[[161   1]
 [ 24 162]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.954898446834
fpr:  [ 0.          0.00617284  0.00617284  0.03703704  0.03703704  0.04320988
  1.        ]
tpr:  [ 0.          0.90860215  0.91397849  0.91397849  0.91935484  0.91935484
  1.        ]
threshold:  [ 2.          1.          0.625       0.5         0.33333333  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 4,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
==================================================================
Optimized KNN hyper parameters
------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.96      0.91       162
        Yes       0.96      0.86      0.91       186

avg / total       0.91      0.91      0.91       348

The accuracy score is 90.80%
---------------------------
Confusion Matrix
---------------------------
[[156   6]
 [ 26 160]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.911589008363
fpr:  [ 0.          0.03703704  1.        ]
tpr:  [ 0.          0.86021505  1.        ]
threshold:  [ 2.  1.  0.]
```

```
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

===================================================================
Program ran in 10.088704109191895 seconds
===================================================================
```

Xu Jun Zhu Bellmore data – pre-processing with TF-IDF + stopword removal

```
===================================================================
KNN classifier results
-------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      0.59      0.69       162
        Yes       0.72      0.89      0.79       186

avg / total       0.77      0.75      0.75       348

The accuracy score is 75.29%
---------------------------
Confusion Matrix
---------------------------
[[ 96  66]
 [ 20 166]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.763689765034
fpr:  [ 0.          0.33950617  0.36419753  0.40740741  0.51234568  0.67901235
  1.        ]
tpr:  [ 0.          0.77419355  0.83333333  0.89247312  0.9516129   0.97849462
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
===================================================================
SVC classifier results
-------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.82      0.59      0.68       162
        Yes       0.71      0.89      0.79       186

avg / total       0.76      0.75      0.74       348

The accuracy score is 74.71%
---------------------------
Confusion Matrix
---------------------------
[[ 95  67]
 [ 21 165]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.754712597903
fpr:  [ 0.          0.          0.         ...,  0.95061728  0.96296296  1.        ]
tpr:  [ 0.00537634  0.01612903  0.03225806 ...,  1.          1.          1.        ]
threshold:  [ 0.95979664  0.91945263  0.90029805 ...,  0.09851521  0.09596514
  0.03168424]
===================================================================
Decision Tree classifier results
-------------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.78      0.64      0.70       162
        Yes       0.73      0.84      0.78       186

avg / total       0.75      0.75      0.74       348

The accuracy score is 74.71%
---------------------------
Confusion Matrix
---------------------------
[[103  59]
 [ 29 157]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.830080977034
fpr:  [ 0.          0.01234568  0.02469136 ...,  0.48765432  0.51234568  1.        ]
tpr:  [ 0.37096774  0.37096774  0.39247312 ...,  0.94623656  0.94623656  1.        ]
threshold:  [ 1.          0.75        0.66666667 ...,  0.16666667  0.09090909  0.        ]
================================================================================
Random Forest classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.78      0.63      0.70       162
        Yes       0.72      0.84      0.78       186

avg / total       0.75      0.74      0.74       348

The accuracy score is 74.43%
---------------------------
Confusion Matrix
---------------------------
[[102  60]
 [ 29 157]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.82384176291
fpr:  [ 0.          0.          0.         ...,  0.67901235  0.69135802  1.        ]
tpr:  [ 0.28494624  0.30107527  0.32795699 ...,  0.97311828  0.97311828  1.        ]
threshold:  [ 1.          0.93333333  0.9        ...,  0.02727273  0.01666667  0.        ]
================================================================================
Bernoulli Naive Bayes classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.74      0.65      0.69       162
        Yes       0.72      0.81      0.76       186

avg / total       0.73      0.73      0.73       348

The accuracy score is 73.28%
---------------------------
Confusion Matrix
---------------------------
[[105  57]
 [ 36 150]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.818382450551
fpr:  [ 0.          0.          0.         ...,  0.96296296  0.98765432  1.        ]
tpr:  [ 0.01075269  0.04301075  0.05376344 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   1.00000000e+00   9.99999996e-01 ...,   1.01006268e-09
   1.60513085e-12   2.41064195e-14]
==================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      0.62      0.72       162
        Yes       0.73      0.91      0.81       186

avg / total       0.79      0.78      0.77       348

The accuracy score is 77.59%
---------------------------
Confusion Matrix
---------------------------
[[100  62]
 [ 16 170]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.789658834462
fpr:  [ 0.          0.          0.         ...,  0.9691358   0.98148148  1.        ]
tpr:  [ 0.01612903  0.02150538  0.03225806 ...,  1.          1.          1.        ]
threshold:  [ 0.88055375  0.82781417  0.79410985 ...,  0.12325863  0.12017839
  0.09242067]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.62      0.70       162
        Yes       0.73      0.87      0.79       186

avg / total       0.76      0.75      0.75       348

The accuracy score is 75.29%
---------------------------
Confusion Matrix
---------------------------
[[101  61]
 [ 25 161]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.834760387628
fpr:  [ 0.          0.          0.         ...,  0.88271605  0.88271605  1.        ]
tpr:  [ 0.1344086   0.13978495  0.19892473 ...,  0.99462366  1.          1.        ]
threshold:  [ 1.          0.99206349  0.98888889 ...,  0.00740741  0.00277778  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 90, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
==================================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.76      0.64      0.69       162
        Yes       0.72      0.83      0.77       186

avg / total       0.74      0.74      0.74       348

The accuracy score is 73.85%
---------------------------
Confusion Matrix
---------------------------
[[103  59]
 [ 32 154]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.835772600558
fpr:  [ 0.          0.00617284  0.01234568 ...,  0.48765432  0.51234568  1.        ]
tpr:  [ 0.37634409  0.37634409  0.38172043 ...,  0.95698925  0.95698925  1.        ]
threshold:  [ 1.          0.77777778  0.75       ...,  0.16666667  0.09090909  0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 4,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
==================================================================
Optimized KNN hyper parameters
------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.58      0.96      0.73       162
        Yes       0.92      0.41      0.57       186

avg / total       0.76      0.66      0.64       348

The accuracy score is 66.38%
---------------------------
Confusion Matrix
---------------------------
[[155   7]
 [110  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.682696136997
fpr:  [ 0.          0.04320988  1.        ]
tpr:  [ 0.          0.40860215  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

==================================================================
Program ran in 5.68264102935791 seconds
==================================================================
```

Bayzick Data – no pre processing

```
=================================================================
KNN classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.85      0.90      0.87        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 94.85%
--------------------------
Confusion Matrix
--------------------------
[[314  13]
 [  8  73]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.976724430853
fpr:  [ 0.          0.0030581   0.00917431  0.03975535  0.08562691  0.11620795
  1.        ]
tpr:  [ 0.          0.64197531  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      1.00      0.91       327
        Yes       0.95      0.22      0.36        81

avg / total       0.86      0.84      0.80       408

The accuracy score is 84.31%
--------------------------
Confusion Matrix
--------------------------
[[326   1]
 [ 63  18]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.9778004304
fpr:  [ 0.          0.          0.0030581  ...,  0.88379205  0.88990826  1.        ]
tpr:  [ 0.01234568  0.20987654  0.20987654 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.99999999e-01   9.99999999e-01 ...,   4.60947279e-03
   4.57952747e-03   4.31219364e-04]
=================================================================
Decision Tree classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.90      0.94      0.92        81

avg / total       0.97      0.97      0.97       408
```

```
The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[319   8]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.956903386567
fpr:  [ 0.          0.02446483  1.        ]
tpr:  [ 0.         0.9382716  1.        ]
threshold:  [ 2.  1.  0.]
===============================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.96      0.99      0.98       327
        Yes       0.97      0.83      0.89        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[325   2]
 [ 14  67]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.99152414392
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.11926606  0.26299694  1.        ]
tpr:  [ 0.          0.32098765  0.55555556 ...,  0.98765432  1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
===============================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.80      0.90      0.85        81

avg / total       0.94      0.94      0.94       408

The accuracy score is 93.63%
---------------------------
Confusion Matrix
---------------------------
[[309  18]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.959980367728
fpr:  [ 0.          0.04281346  0.04281346 ...,  0.3088685   0.31804281  1.        ]
tpr:  [ 0.          0.66666667  0.72839506 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   1.45815273e-023   1.36144112e-023   1.01730288e-133]
=============================================================
Optimized SVC hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.98      0.98       327
        Yes       0.93      0.95      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
--------------------------
Confusion Matrix
--------------------------
[[321   6]
 [  4  77]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.991354249254
fpr:  [ 0.          0.          0.0030581  ...,  0.68195719  0.6911315   1.        ]
tpr:  [ 0.01234568  0.77777778  0.77777778 ...,  1.          1.          1.        ]
threshold:  [  9.99999883e-01   9.43362714e-01   9.35292316e-01 ...,   4.17740877e-03
   4.09364703e-03   2.56464223e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
=============================================================
Optimized Random Forest hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.98       327
        Yes       0.97      0.90      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
--------------------------
Confusion Matrix
--------------------------
[[325   2]
 [  8  73]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.991373126439
fpr:  [ 0.          0.          0.         ...,  0.41896024  0.59633028  1.        ]
tpr:  [ 0.01234568  0.09876543  0.19753086 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=============================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.96      0.95      0.96       327
        Yes       0.82      0.84      0.83        81

avg / total       0.93      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[312  15]
 [ 13  68]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.859893532676
fpr:  [ 0.          0.07033639  1.        ]
tpr:  [ 0.          0.79012346  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96149054253
fpr:  [ 0.          0.01529052  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 44.42751216888428 seconds
=================================================================
```

Bayzick data – pre-processing with porter stemmer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
```

```
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.977309623589
fpr:  [ 0.          0.0030581   0.00917431  0.03669725  0.0795107   0.11314985
  1.        ]
tpr:  [ 0.          0.65432099  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
====================================================================
SVC classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.87      1.00      0.93       327
        Yes       0.97      0.41      0.57        81

avg / total       0.89      0.88      0.86       408

The accuracy score is 87.99%
---------------------------
Confusion Matrix
---------------------------
[[326   1]
 [ 48  33]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.981840147997
fpr:  [ 0.          0.          0.0030581  ...,  0.60244648  0.60244648  1.        ]
tpr:  [ 0.01234568  0.27160494  0.27160494 ...,  0.98765432  1.          1.        ]
threshold:  [  1.00000000e+00   9.99999993e-01   9.99999988e-01 ...,   1.25021210e-02
   1.24739295e-02   1.35873090e-04]
====================================================================
Decision Tree classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.95      0.96       327
        Yes       0.83      0.88      0.85        81

avg / total       0.94      0.94      0.94       408

The accuracy score is 93.87%
---------------------------
Confusion Matrix
---------------------------
[[312  15]
 [ 10  71]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.915335825122
fpr:  [ 0.          0.04587156  1.        ]
tpr:  [ 0.          0.87654321  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.92      0.88      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.988371654019
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.11926606  0.32110092  1.        ]
tpr:  [ 0.          0.2962963   0.5308642  ...,  0.98765432  1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.79      0.88      0.83        81

avg / total       0.93      0.93      0.93       408

The accuracy score is 92.89%
---------------------------
Confusion Matrix
---------------------------
[[308  19]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.961584928455
fpr:  [ 0.          0.03363914  0.03363914 ...,  0.37308869  0.382263    1.        ]
tpr:  [ 0.          0.61728395  0.64197531 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   7.00679586e-021   5.98519633e-021   3.27734584e-128]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.93      0.94      0.93        81
```

```
avg / total       0.97      0.97      0.97       408

The accuracy score is 97.30%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.996111299883
fpr:  [ 0.          0.          0.0030581  ...,  0.79204893  0.80122324  1.        ]
tpr:  [ 0.01234568  0.56790123  0.56790123 ...,  1.          1.          1.        ]
threshold:  [  9.99992368e-01   9.55965489e-01   9.55792623e-01 ...,   2.24463361e-03
   2.22432802e-03   1.54933558e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
============================================================
Optimized Random Forest hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.95      0.89      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993676143014
fpr:  [ 0.          0.          0.         ...,  0.40978593  0.62079511  1.        ]
tpr:  [ 0.04938272  0.13580247  0.19753086 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
============================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.95       327
        Yes       0.85      0.75      0.80        81

avg / total       0.92      0.92      0.92       408

The accuracy score is 92.40%
---------------------------
Confusion Matrix
```

```
---------------------------
[[316  11]
 [ 20  61]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.884471627591
fpr:  [ 0.          0.04587156  1.        ]
tpr:  [ 0.          0.81481481  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.93      0.93      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.06%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953788651036
fpr:  [ 0.          0.01834862  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 64.9511501789093 seconds
================================================================
```

Bayzick data – pre-processing with lemmatizer

```
================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.85      0.90      0.87        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 94.85%
---------------------------
Confusion Matrix
---------------------------
[[314  13]
```

```
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.976894325518
fpr:  [ 0.          0.0030581   0.00917431  0.03975535  0.08256881  0.11620795
  1.        ]
tpr:  [ 0.          0.65432099  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.86      1.00      0.92       327
        Yes       0.96      0.32      0.48        81

avg / total       0.88      0.86      0.83       408

The accuracy score is 86.27%
---------------------------
Confusion Matrix
---------------------------
[[326   1]
 [ 55  26]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.977158606109
fpr:  [ 0.          0.          0.0030581  ...,  0.64831804  0.65443425  1.        ]
tpr:  [ 0.01234568  0.25925926  0.25925926 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.99999997e-01   9.99999997e-01 ...,   1.17812572e-02
   1.17364990e-02   3.55846565e-04]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.84      0.93      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[313  14]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.941556235134
fpr:  [ 0.          0.04281346  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.93      0.86      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [ 11  70]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.9916751614
fpr:  [ 0.          0.0030581   0.00611621 ...,  0.11314985  0.25382263  1.        ]
tpr:  [ 0.24691358  0.56790123  0.67901235 ...,  1.          1.          1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
==================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.79      0.89      0.84        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[308  19]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.959980367728
fpr:  [ 0.          0.04281346  0.04281346 ...,  0.35168196  0.36085627  1.        ]
tpr:  [ 0.          0.66666667  0.67901235 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   1.08123869e-023   9.40106890e-024   7.44845718e-135]
==================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.98      0.98       327
        Yes       0.93      0.95      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  4  77]]
```

```
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.99437459886
fpr:  [ 0.          0.          0.0030581  ...,  0.65137615  0.66055046  1.        ]
tpr:  [ 0.01234568  0.64197531  0.64197531 ...,  1.          1.          1.        ]
threshold:  [  9.99995580e-01   9.47990586e-01   9.47591672e-01 ...,   5.73964294e-03
   5.70170978e-03   4.85652641e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
===============================================================
Optimized Random Forest hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.92      0.90      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993846037679
fpr:  [ 0.          0.          0.         ...,  0.40366972  0.62079511  1.        ]
tpr:  [ 0.0617284   0.08641975  0.12345679 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
===============================================================
Optimized Decision Tree hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.94      0.94       327
        Yes       0.77      0.79      0.78        81

avg / total       0.91      0.91      0.91       408

The accuracy score is 91.18%
---------------------------
Confusion Matrix
---------------------------
[[308  19]
 [ 17  64]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.881300260505
fpr:  [ 0.          0.02752294  1.        ]
tpr:  [ 0.          0.79012346  1.        ]
```

```
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96149054253
fpr:  [ 0.          0.01529052  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 55.730018615722656 seconds
================================================================
```

Bayzick data – pre-processing with bi-grams

```
================================================================
KNN classifier results
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.977139728924
fpr:  [ 0.          0.0030581   0.00917431  0.03669725  0.08562691  0.11009174
  1.        ]
tpr:  [ 0.          0.65432099  0.80246914  0.90123457  0.96296296  0.97530864
```

```
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
================================================================
SVC classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      1.00      0.91       327
        Yes       1.00      0.22      0.36        81

avg / total       0.87      0.85      0.80       408

The accuracy score is 84.56%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 63  18]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.979272850832
fpr:  [ 0.          0.          0.0030581  ...,  0.7675841   0.77370031  1.        ]
tpr:  [ 0.01234568  0.22222222  0.22222222 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.99999996e-01   9.99999993e-01 ...,   8.79704278e-03
   8.75032431e-03   5.29270764e-04]
================================================================
Decision Tree classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.97      0.98       327
        Yes       0.89      0.91      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[318   9]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.943028655567
fpr:  [ 0.          0.02752294  1.        ]
tpr:  [ 0.          0.91358025  1.        ]
threshold:  [ 2.  1.  0.]
================================================================
Random Forest classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.97       327
        Yes       0.91      0.88      0.89        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 95.83%
```

```
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.990636916223
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.13455657  0.3058104   1.        ]
tpr:  [ 0.          0.24691358  0.4691358  ...,  1.          1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.80      0.89      0.84        81

avg / total       0.94      0.93      0.94       408

The accuracy score is 93.38%
---------------------------
Confusion Matrix
---------------------------
[[309  18]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96133952505
fpr:  [ 0.          0.03975535  0.04281346 ...,  0.32415902  0.33333333  1.        ]
tpr:  [ 0.          0.66666667  0.66666667 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   6.40458255e-024   6.35883693e-024   9.34615827e-134]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.98      0.98       327
        Yes       0.93      0.95      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  4  77]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.991769547325
fpr:  [ 0.          0.          0.0030581  ...,  0.77981651  0.78899083  1.        ]
tpr:  [ 0.01234568  0.65432099  0.65432099 ...,  1.          1.          1.        ]
threshold:  [  9.99999912e-01   9.67870621e-01   9.67629361e-01 ...,   2.08296805e-03
   2.06990345e-03   3.56622900e-05]
```

```
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.94      0.89      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993034318722
fpr:  [ 0.          0.          0.         ...,  0.37308869  0.59327217  1.        ]
tpr:  [ 0.0617284   0.13580247  0.30864198 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.94 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.96      0.95      0.96       327
        Yes       0.82      0.85      0.84        81

avg / total       0.93      0.93      0.93       408

The accuracy score is 93.38%
---------------------------
Confusion Matrix
---------------------------
[[312  15]
 [ 12  69]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.893645939517
fpr:  [ 0.          0.02752294  1.        ]
tpr:  [ 0.          0.81481481  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2, 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96149054253
fpr:  [ 0.          0.01529052  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 52.67899250984192 seconds
=================================================================
```

Bayzick data – pre-processing with tri-grams

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.977139728924
fpr:  [ 0.          0.0030581   0.00917431  0.03669725  0.08562691  0.11009174
  1.        ]
tpr:  [ 0.          0.65432099  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      1.00      0.91       327
        Yes       1.00      0.22      0.36        81

avg / total       0.87      0.85      0.80       408

The accuracy score is 84.56%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 63  18]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.979272850832
fpr:  [ 0.          0.          0.0030581  ...,  0.7675841   0.77370031  1.        ]
tpr:  [ 0.01234568  0.22222222  0.22222222 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.99999998e-01   9.99999997e-01 ...,   7.36038162e-03
   7.31983268e-03   3.98790151e-04]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.94      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 95.83%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950787178616
fpr:  [ 0.          0.03669725  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      1.00      0.99       327
        Yes       0.99      0.90      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.79%
---------------------------
Confusion Matrix
---------------------------
[[326   1]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.989447653566
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.11926606  0.33944954  1.        ]
tpr:  [ 0.          0.33333333  0.59259259 ...,  0.97530864  1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.80      0.89      0.84        81

avg / total       0.94      0.93      0.94       408

The accuracy score is 93.38%
---------------------------
Confusion Matrix
---------------------------
[[309  18]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96133952505
fpr:  [ 0.          0.03975535  0.04281346 ...,  0.32415902  0.33333333  1.        ]
tpr:  [ 0.          0.66666667  0.66666667 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   6.40458255e-024   6.35883693e-024   9.34615827e-134]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.98      0.98       327
        Yes       0.93      0.95      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  4  77]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.991769547325
fpr:  [ 0.          0.          0.0030581  ...,  0.77981651  0.78899083  1.        ]
tpr:  [ 0.01234568  0.65432099  0.65432099 ...,  1.          1.          1.        ]
threshold:  [  9.99999910e-01   9.68399134e-01   9.68163111e-01 ...,   2.24199194e-03
   2.22801260e-03   3.93158902e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.98      0.98      0.98       327
        Yes       0.93      0.91      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993147581833
fpr:  [ 0.          0.          0.         ...,  0.3883792  0.617737   1.        ]
tpr:  [ 0.02469136  0.04938272  0.12345679 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.94      0.94       327
        Yes       0.75      0.79      0.77        81

avg / total       0.91      0.91      0.91       408

The accuracy score is 90.69%
---------------------------
Confusion Matrix
---------------------------
[[306  21]
 [ 17  64]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.867538792615
fpr:  [ 0.          0.05504587  1.        ]
tpr:  [ 0.          0.79012346  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2, 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
```

```
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96149054253
fpr:  [ 0.          0.01529052  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 54.75537419319153 seconds
================================================================
```

Bayzick data – pre-processing with 4-grams

```
================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.977139728924
fpr:  [ 0.          0.0030581   0.00917431  0.03669725  0.08562691  0.11009174
  1.        ]
tpr:  [ 0.          0.65432099  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.84      1.00      0.91       327
        Yes       1.00      0.22      0.36        81

avg / total       0.87      0.85      0.80       408

The accuracy score is 84.56%
---------------------------
Confusion Matrix
---------------------------
```

```
[[327   0]
 [ 63  18]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.979272850832
fpr:  [ 0.          0.          0.0030581  ...,  0.7675841   0.77370031  1.        ]
tpr:  [ 0.01234568  0.22222222  0.22222222 ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.99999995e-01   9.99999991e-01 ...,   8.74724164e-03
   8.70112026e-03   5.37012905e-04]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.97      0.98       327
        Yes       0.88      0.93      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[317  10]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.947672443085
fpr:  [ 0.          0.03058104  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.95      0.88      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.991467512365
fpr:  [ 0.          0.          0.0030581  ...,  0.08256881  0.25382263  1.        ]
tpr:  [ 0.17283951  0.4691358   0.7037037  ...,  0.98765432  1.          1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.97      0.94      0.96       327
        Yes       0.80      0.89      0.84        81

avg / total       0.94      0.93      0.94       408

The accuracy score is 93.38%
---------------------------
Confusion Matrix
---------------------------
[[309  18]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96133952505
fpr:  [ 0.          0.03975535  0.04281346 ...,  0.32415902  0.33333333  1.        ]
tpr:  [ 0.          0.66666667  0.66666667 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   6.40458255e-024   6.35883693e-024   9.34615827e-134]
==================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.98      0.98       327
        Yes       0.93      0.95      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  4  77]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.991769547325
fpr:  [ 0.          0.          0.0030581  ...,  0.77981651  0.78899083  1.        ]
tpr:  [ 0.01234568  0.65432099  0.65432099 ...,  1.          1.          1.        ]
threshold:  [  9.99999896e-01   9.65534108e-01   9.65276611e-01 ...,   1.98469441e-03
   1.97227770e-03   3.43330012e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.97      0.89      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.30%
---------------------------
Confusion Matrix
```

```
---------------------------
[[325   2]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.991920564805
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.37920489  0.60244648  1.        ]
tpr:  [ 0.          0.01234568  0.0617284  ...,  1.          1.          1.        ]
threshold:  [ 2.    1.    0.98 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.96      0.94      0.95       327
        Yes       0.79      0.83      0.81        81

avg / total       0.92      0.92      0.92       408

The accuracy score is 92.16%
---------------------------
Confusion Matrix
---------------------------
[[309  18]
 [ 14  67]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.856722165591
fpr:  [ 0.          0.05198777  1.        ]
tpr:  [ 0.         0.7654321  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.96149054253
```

```
fpr:  [ 0.          0.01529052  1.        ]
tpr:  [ 0.         0.9382716  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 59.147298097610474 seconds
================================================================
```

Bayzick data – pre-processing with stop word removal

```
================================================================
KNN classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.976120360932
fpr:  [ 0.          0.0030581   0.01834862  0.03669725  0.08868502  0.11009174
  1.        ]
tpr:  [ 0.          0.67901235  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
================================================================
SVC classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.83      1.00      0.91       327
        Yes       1.00      0.17      0.29        81

avg / total       0.86      0.84      0.79       408

The accuracy score is 83.58%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 67  14]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978744289652
fpr:  [ 0.          0.          0.0030581  ...,  0.82568807  0.82568807  1.        ]
tpr:  [ 0.02469136  0.20987654  0.20987654 ...,  0.98765432  1.          1.        ]
```

```
threshold:  [ 1.          1.          0.99999999 ...,  0.01316939  0.01299141
  0.00384562]
================================================================
Decision Tree classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.93      0.91      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
--------------------------
Confusion Matrix
--------------------------
[[321   6]
 [  7  74]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.94761581153
fpr:  [ 0.          0.01834862  1.        ]
tpr:  [ 0.          0.91358025  1.        ]
threshold:  [ 2.  1.  0.]
================================================================
Random Forest classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.97      0.86      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
--------------------------
Confusion Matrix
--------------------------
[[325   2]
 [ 11  70]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.98916449579
fpr:  [ 0.          0.          0.0030581  ...,  0.05810398  0.12232416  1.        ]
tpr:  [ 0.24691358  0.60493827  0.75308642 ...,  0.97530864  0.98765432  1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
================================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.96       327
        Yes       0.78      0.90      0.84        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
--------------------------
```

```
Confusion Matrix
---------------------------
[[307  20]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.97236380111
fpr:  [ 0.          0.03058104  0.03058104 ...,  0.22018349  0.22629969  1.        ]
tpr:  [ 0.          0.75308642  0.77777778 ...,  1.          1.          1.        ]
threshold:  [  2.00000000e+00   1.00000000e+00   1.00000000e+00 ...,   2.31052935e-18
   2.21401283e-18   7.22079152e-88]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.94      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.99426133575
fpr:  [ 0.          0.          0.0030581  ...,  0.44648318  0.45259939  1.        ]
tpr:  [ 0.01234568  0.69135802  0.69135802 ...,  1.          1.          1.        ]
threshold:  [  9.99999630e-01   9.72080021e-01   9.70293093e-01 ...,   1.29348386e-02
   1.29185100e-02   1.21983118e-05]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.99       327
        Yes       0.97      0.91      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.79%
---------------------------
Confusion Matrix
---------------------------
[[325   2]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.988994601125
fpr:  [ 0.          0.          0.         ...,  0.12844037  0.2324159   1.        ]
```

```
tpr:  [ 0.07407407  0.2345679   0.30864198 ...,  0.98765432  0.98765432  1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.97      0.97       327
        Yes       0.87      0.88      0.87        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 94.85%
---------------------------
Confusion Matrix
---------------------------
[[316  11]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.932268660097
fpr:  [ 0.          0.03669725  1.        ]
tpr:  [ 0.          0.90123457  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.99       327
        Yes       0.97      0.93      0.95        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 98.04%
---------------------------
Confusion Matrix
---------------------------
[[325   2]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.959904858987
fpr:  [ 0.          0.00611621  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 39.01992082595825 seconds
```

```
=================================================================
```

Bayzick data – pre-processing with TF-IDF

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978253482841
fpr:  [ 0.          0.0030581   0.00611621  0.03669725  0.0795107   0.09785933
  1.        ]
tpr:  [ 0.          0.67901235  0.81481481  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      1.00      0.89       327
        Yes       0.00      0.00      0.00        81

avg / total       0.64      0.80      0.71       408

The accuracy score is 80.15%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 81   0]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.988409408389
fpr:  [ 0.          0.          0.0030581  ...,  0.72477064  0.73088685  1.        ]
tpr:  [ 0.01234568  0.20987654  0.20987654 ...,  1.          1.          1.        ]
threshold:  [ 1.          1.          1.         ...,  0.01406506  0.01402633
  0.00571137]
C:\Users\peter\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.98      0.97      0.98       327
        Yes       0.89      0.93      0.91        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.32%
---------------------------
Confusion Matrix
---------------------------
[[318   9]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.949201495073
fpr:  [ 0.          0.02752294  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.94      0.89      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987956355948
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.08868502  0.22018349  1.        ]
tpr:  [ 0.          0.43209877  0.66666667 ...,  0.97530864  1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.92      0.96       327
        Yes       0.76      0.96      0.85        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[302  25]
 [  3  78]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.965379242647
fpr:  [ 0.          0.05504587  0.05504587 ...,  0.24159021  0.24159021  1.        ]
tpr:  [ 0.          0.90123457  0.92592593 ...,  0.98765432  1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   3.85636094e-060   3.06036509e-060   1.12120149e-304]
==================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.93      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.30%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.98874919772
fpr:  [ 0.          0.          0.0030581  ...,  0.71559633  0.71559633  1.        ]
tpr:  [ 0.01234568  0.77777778  0.77777778 ...,  0.98765432  1.          1.        ]
threshold:  [  9.99999998e-01   9.43007472e-01   9.41011094e-01 ...,   1.31478163e-02
   1.30678710e-02   4.25754372e-06]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.94      0.89      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.983765620871
fpr:  [ 0.          0.          0.         ...,  0.21712538  0.44648318  1.        ]
tpr:  [ 0.09876543  0.32098765  0.38271605 ...,  0.98765432  0.98765432  1.        ]
threshold:  [ 1.    0.96  0.94 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.97      0.97       327
        Yes       0.88      0.89      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.34%
---------------------------
Confusion Matrix
---------------------------
[[317  10]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.904405934987
fpr:  [ 0.          0.01834862  1.        ]
tpr:  [ 0.          0.82716049  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.93      0.93      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.06%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953788651036
fpr:  [ 0.          0.01834862  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 70.01957368850708 seconds
=================================================================
```

Bayzick data – pre-processing with TF-IDF + porter stemmer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.97764941292
fpr:  [ 0.          0.0030581   0.00611621  0.03669725  0.0764526   0.11009174
  1.        ]
tpr:  [ 0.          0.65432099  0.79012346  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      1.00      0.89       327
        Yes       0.00      0.00      0.00        81

avg / total       0.64      0.80      0.71       408

The accuracy score is 80.15%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 81   0]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987956355948
fpr:  [ 0.          0.          0.0030581  ...,  0.83792049  0.8470948   1.        ]
tpr:  [ 0.01234568  0.2345679   0.2345679  ...,  1.          1.          1.        ]
threshold:  [ 1.          0.99999998  0.99999998 ...,  0.01794134  0.01785402
  0.00839019]
C:\Users\peter\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support
        No       0.98      0.98      0.98       327
       Yes       0.90      0.91      0.91        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.32%
---------------------------
Confusion Matrix
---------------------------
[[319   8]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.944557707555
fpr:  [ 0.          0.02446483  1.        ]
tpr:  [ 0.          0.91358025  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

        No       0.97      0.98      0.97       327
       Yes       0.91      0.88      0.89        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 95.83%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987050251067
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.0764526   0.20183486  1.        ]
tpr:  [ 0.          0.40740741  0.62962963 ...,  0.96296296  1.          1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

        No       0.99      0.92      0.96       327
       Yes       0.76      0.96      0.85        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[302  25]
 [  3  78]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.965454751387
fpr:  [ 0.          0.05504587  0.05504587 ...,  0.2293578   0.2293578   1.        ]
tpr:  [ 0.          0.90123457  0.92592593 ...,  0.98765432  1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   2.39572285e-058   1.59816304e-058   2.56823103e-308]
================================================================
Optimized SVC hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.93      0.93      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.06%
---------------------------
Confusion Matrix
---------------------------
[[321   6]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.98916449579
fpr:  [ 0.          0.          0.0030581  ...,  0.82262997  0.82874618  1.        ]
tpr:  [ 0.01234568  0.7654321   0.7654321  ...,  1.          1.          1.        ]
threshold:  [  1.00000000e+00   9.66289685e-01   9.64101486e-01 ...,   9.93817419e-03
   9.87123436e-03   3.05913403e-06]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
================================================================
Optimized Random Forest hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.94      0.89      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993921546419
fpr:  [ 0.          0.          0.         ...,  0.2293578   0.43119266  1.        ]
tpr:  [ 0.03703704  0.16049383  0.24691358 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.97      0.97       327
        Yes       0.87      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.34%
---------------------------
Confusion Matrix
---------------------------
[[316  11]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.912164458036
fpr:  [ 0.          0.02752294  1.        ]
tpr:  [ 0.          0.85185185  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
======================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.91      0.93      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.952259599049
fpr:  [ 0.          0.02140673  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

======================================================================
Program ran in 93.48810338973999 seconds
======================================================================
```

Bayzick data – pre-processing with TF-IDF + lemmatizer

```
======================================================================
KNN classifier results
```

```
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.97783818477
fpr:  [ 0.          0.0030581   0.00611621  0.03669725  0.08256881  0.10091743
  1.        ]
tpr:  [ 0.          0.66666667  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      1.00      0.89       327
        Yes       0.00      0.00      0.00        81

avg / total       0.64      0.80      0.71       408

The accuracy score is 80.15%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 81   0]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987729829728
fpr:  [ 0.          0.          0.0030581  ...,  0.82262997  0.83180428  1.        ]
tpr:  [ 0.01234568  0.19753086  0.19753086 ...,  1.          1.          1.        ]
threshold:  [ 1.          1.          1.         ...,  0.01393185  0.01392925
  0.00613718]
C:\Users\peter\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.95      0.96       327
        Yes       0.83      0.89      0.86        81

avg / total       0.94      0.94      0.94       408
```

```
The accuracy score is 94.12%
---------------------------
Confusion Matrix
---------------------------
[[312  15]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.921508664628
fpr:  [ 0.          0.04587156  1.        ]
tpr:  [ 0.          0.88888889  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.96      0.99      0.98       327
        Yes       0.95      0.85      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [ 12  69]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987673198173
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.05504587  0.16819572  1.        ]
tpr:  [ 0.2962963   0.59259259  0.7654321  ...,  0.97530864  0.98765432  1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.92      0.96       327
        Yes       0.76      0.96      0.85        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[302  25]
 [  3  78]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965492505758
fpr:  [ 0.          0.05504587  0.05810398 ...,  0.26299694  0.26299694  1.        ]
tpr:  [ 0.          0.91358025  0.91358025 ...,  0.98765432  1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   8.89533080e-059   5.32623156e-059   1.19056910e-310]
```

```
================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.94      0.93      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.30%
---------------------------
Confusion Matrix
---------------------------
[[322   5]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987616566618
fpr:  [ 0.          0.          0.0030581  ...,  0.78899083  0.78899083  1.        ]
tpr:  [ 0.01234568  0.74074074  0.74074074 ...,  0.98765432  1.          1.        ]
threshold:  [  9.99999999e-01   9.55701494e-01   9.54242647e-01 ...,   1.04482627e-02
   1.03726176e-02   3.12070403e-06]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.91      0.90      0.91        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.32%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.993487371163
fpr:  [ 0.          0.          0.         ...,  0.25076453  0.44648318  1.        ]
tpr:  [ 0.02469136  0.11111111  0.25925926 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.97      0.97       327
```

```
       Yes       0.87      0.88      0.87        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 94.85%
---------------------------
Confusion Matrix
---------------------------
[[316  11]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.918450560652
fpr:  [ 0.          0.05198777  1.        ]
tpr:  [ 0.          0.88888889  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.91      0.93      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.952259599049
fpr:  [ 0.          0.02140673  1.        ]
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 81.45110630989075 seconds
=================================================================
```

Bayzick data – pre-processing with TF-IDF + tri-grams

```
=================================================================
KNN classifier results
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
```

```
        Yes       0.86      0.90      0.88        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[315  12]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.97798920225
fpr:  [ 0.          0.0030581   0.00611621  0.03669725  0.0795107   0.10091743
  1.        ]
tpr:  [ 0.          0.67901235  0.80246914  0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      1.00      0.89       327
        Yes       0.00      0.00      0.00        81

avg / total       0.64      0.80      0.71       408

The accuracy score is 80.15%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 81   0]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.986370672405
fpr:  [ 0.          0.          0.0030581  ...,  0.20183486  0.20183486  1.        ]
tpr:  [ 0.01234568  0.20987654  0.20987654 ...,  0.98765432  1.          1.        ]
threshold:  [ 1.          1.          0.99999999 ...,  0.02431839  0.02430163
  0.0075061 ]
C:\Users\peter\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.96      0.97       327
        Yes       0.86      0.91      0.89        81

avg / total       0.95      0.95      0.95       408

The accuracy score is 95.34%
---------------------------
Confusion Matrix
---------------------------
```

```
[[315  12]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.938441499604
fpr:  [ 0.          0.03669725  1.        ]
tpr:  [ 0.          0.91358025  1.        ]
threshold:  [ 2.  1.  0.]
================================================================
Random Forest classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.98      0.98       327
        Yes       0.91      0.89      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.976799939593
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.09480122  0.2293578   1.        ]
tpr:  [ 0.          0.37037037  0.58024691 ...,  0.97530864  0.97530864  1.        ]
threshold:  [ 2.   1.   0.9 ...,  0.2  0.1  0. ]
================================================================
Bernoulli Naive Bayes classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.92      0.96       327
        Yes       0.76      0.96      0.85        81

avg / total       0.94      0.93      0.93       408

The accuracy score is 93.14%
---------------------------
Confusion Matrix
---------------------------
[[302  25]
 [  3  78]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.9647374183556
fpr:  [ 0.          0.05810398  0.07033639 ...,  0.25993884  0.98165138  1.        ]
tpr:  [ 0.         0.9382716  0.9382716 ...,  1.         1.         1.       ]
threshold:  [  2.00000000e+000   1.00000000e+000   9.99974087e-001 ...,
   5.20410960e-087   7.25839894e-316   0.00000000e+000]
================================================================
Optimized SVC hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.98      0.98      0.98       327
        Yes       0.92      0.94      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.06%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.989579793861
fpr:  [ 0.          0.          0.0030581  ...,  0.65443425  0.65443425  1.        ]
tpr:  [ 0.01234568  0.81481481  0.81481481 ...,  0.98765432  1.          1.        ]
threshold:  [  9.99999988e-01   9.56593035e-01   9.51049513e-01 ...,   1.13041628e-02
   1.12230127e-02   1.19104300e-06]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
================================================================
Optimized Random Forest hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.98       327
        Yes       0.96      0.90      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.30%
---------------------------
Confusion Matrix
---------------------------
[[324   3]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.992864424057
fpr:  [ 0.          0.          0.         ...,  0.23547401  0.46788991  1.        ]
tpr:  [ 0.04938272  0.17283951  0.27160494 ...,  1.          1.          1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.94      0.95       327
        Yes       0.78      0.86      0.82        81

avg / total       0.93      0.92      0.93       408
```

```
The accuracy score is 92.40%
---------------------------
Confusion Matrix
---------------------------
[[307  20]
 [ 11  70]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.932212028542
fpr:  [ 0.          0.02446483  1.        ]
tpr:  [ 0.          0.88888889  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.92      0.94      0.93        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 97.06%
---------------------------
Confusion Matrix
---------------------------
[[320   7]
 [  5  76]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.958432438555
fpr:  [ 0.          0.02140673  1.        ]
tpr:  [ 0.         0.9382716  1.       ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 102.95309567451477 seconds
=================================================================
```

Bayzick data – pre-processing with TF-IDF + stop word removal

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.97      0.97       327
        Yes       0.89      0.90      0.90        81

avg / total       0.96      0.96      0.96       408
```

```
The accuracy score is 95.83%
---------------------------
Confusion Matrix
---------------------------
[[318   9]
 [  8  73]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.978517763431
fpr:  [ 0.          0.0030581   0.00611621  0.02752294  0.0733945   0.08562691
  1.        ]
tpr:  [ 0.          0.59259259  0.7654321   0.90123457  0.96296296  0.97530864
  1.        ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.81      1.00      0.89       327
        Yes       1.00      0.05      0.09        81

avg / total       0.85      0.81      0.74       408

The accuracy score is 81.13%
---------------------------
Confusion Matrix
---------------------------
[[327   0]
 [ 77   4]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.986181900555
fpr:  [ 0.          0.          0.0030581  ...,  0.89602446  0.90519878  1.        ]
tpr:  [ 0.01234568  0.17283951  0.17283951 ...,  1.          1.          1.        ]
threshold:  [ 1.          1.          1.         ...,  0.01101245  0.01091287
  0.00290242]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.97      0.98       327
        Yes       0.88      0.93      0.90        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.08%
---------------------------
Confusion Matrix
---------------------------
[[317  10]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.947672443085
fpr:  [ 0.          0.03058104  1.        ]
```

```
tpr:  [ 0.          0.92592593  1.        ]
threshold:  [ 2.  1.  0.]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.95      0.88      0.91        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.57%
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [ 10  71]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.98897572394
fpr:  [ 0.          0.0030581   0.0030581  ...,  0.03975535  0.0795107   1.        ]
tpr:  [ 0.44444444  0.64197531  0.7654321  ...,  0.97530864  0.98765432  1.        ]
threshold:  [ 1.   0.9  0.8 ...,  0.2  0.1  0. ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.94      0.96       327
        Yes       0.79      0.95      0.86        81

avg / total       0.95      0.94      0.94       408

The accuracy score is 93.87%
---------------------------
Confusion Matrix
---------------------------
[[306  21]
 [  4  77]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.970362819496
fpr:  [ 0.          0.04281346  0.04281346 ...,  0.33639144  0.33639144  1.        ]
tpr:  [ 0.          0.87654321  0.9382716  ...,  0.98765432  1.          1.        ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   1.71158213e-031   9.89188250e-032   2.87886987e-166]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.98       327
        Yes       0.95      0.93      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.55%
```

```
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [  6  75]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.987616566618
fpr:  [ 0.          0.          0.0030581  ...,  0.90519878  0.91131498  1.        ]
tpr:  [ 0.01234568  0.51851852  0.51851852 ...,  1.          1.          1.        ]
threshold:  [  9.99999824e-01   9.74678657e-01   9.74087655e-01 ...,   6.52152515e-03
   6.11982853e-03   1.61486594e-06]
Parameters were:  {'C': 1000, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape':
None, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None,
'shrinking': True, 'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.97      0.99      0.98       327
        Yes       0.95      0.89      0.92        81

avg / total       0.97      0.97      0.97       408

The accuracy score is 96.81%
---------------------------
Confusion Matrix
---------------------------
[[323   4]
 [  9  72]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.981085060596
fpr:  [ 0.          0.          0.0030581  ...,  0.11009174  0.15902141  1.        ]
tpr:  [ 0.22222222  0.28395062  0.40740741 ...,  0.97530864  0.97530864  1.        ]
threshold:  [ 1.    0.98  0.96 ...,  0.04  0.02  0.  ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features':
'sqrt', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 50, 'n_jobs': 1, 'oob_score': False, 'random_state': None,
'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.98      0.98       327
        Yes       0.90      0.91      0.91        81

avg / total       0.96      0.96      0.96       408

The accuracy score is 96.32%
---------------------------
Confusion Matrix
---------------------------
[[319   8]
 [  7  74]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.933797712085
fpr:  [ 0.          0.03363914  1.        ]
tpr:  [ 0.          0.90123457  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 2, 
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2, 
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.99      0.99       327
        Yes       0.97      0.91      0.94        81

avg / total       0.98      0.98      0.98       408

The accuracy score is 97.79%
---------------------------
Confusion Matrix
---------------------------
[[325   2]
 [  7  74]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953732019481
fpr:  [ 0.          0.00611621  1.        ]
tpr:  [ 0.          0.91358025  1.        ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'manhattan', 'metric_params': None, 'n_jobs': 1, 
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 43.67701244354248 seconds
=================================================================
```

Combined data – no pre-processing

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.95      0.93       505
        Yes       0.89      0.80      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 90.07%
---------------------------
Confusion Matrix
---------------------------
[[480  25]
 [ 50 200]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.949904950495
fpr:  [ 0.          0.00594059  0.02574257  0.04950495  0.12079208  0.21188119
  1.        ]
tpr:  [ 0.     0.604  0.672  0.8    0.904  0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.79      0.96      0.87       505
        Yes       0.87      0.50      0.63       250

avg / total       0.82      0.81      0.79       755

The accuracy score is 80.79%
---------------------------
Confusion Matrix
---------------------------
[[486  19]
 [126 124]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.912946534653
fpr:  [ 0.          0.          0.         ...,  0.99405941  0.9980198   1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.99999988  0.99999714  0.99999472 ...,  0.01608771  0.01230418
  0.0105068 ]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.93       505
        Yes       0.92      0.80      0.85       250

avg / total       0.91      0.91      0.91       755

The accuracy score is 90.99%
---------------------------
Confusion Matrix
---------------------------
[[488  17]
 [ 51 199]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.884158415842
fpr:  [ 0.          0.03366337  0.03366337  0.04158416  1.        ]
tpr:  [ 0.     0.792  0.796  0.804  1.   ]
threshold:  [ 2.          1.          0.66666667  0.5         0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.97      0.93       505
        Yes       0.94      0.77      0.84       250
```

```
avg / total       0.91      0.91      0.90       755

The accuracy score is 90.60%
----------------------------
Confusion Matrix
----------------------------
[[492  13]
 [ 58 192]]
----------------------------
ROC Curve
----------------------------
Area under the curve:  0.96380990099
fpr:  [ 0.          0.0039604   0.01386139 ...,  0.37227723  0.38019802  1.        ]
tpr:  [ 0.32   0.476  0.596 ...,  0.98   0.98   1.   ]
threshold:  [ 1.    0.9   0.8  ...,  0.1   0.05  0.  ]
====================================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.62      0.75       505
        Yes       0.55      0.92      0.68       250

avg / total       0.81      0.72      0.73       755

The accuracy score is 71.92%
----------------------------
Confusion Matrix
----------------------------
[[313 192]
 [ 20 230]]
----------------------------
ROC Curve
----------------------------
Area under the curve:  0.846906930693
fpr:  [ 0.          0.18613861  0.18613861 ...,  0.96039604  0.96039604  1.        ]
tpr:  [ 0.     0.768  0.776 ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   6.12773017e-100   8.74441896e-102   4.94077553e-177]
====================================================================
Optimized SVC hyper parameters
--------------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.95      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.44%
----------------------------
Confusion Matrix
----------------------------
[[493  12]
 [ 30 220]]
----------------------------
ROC Curve
----------------------------
Area under the curve:  0.970697029703
fpr:  [ 0.          0.          0.          ...,  0.87524752  0.87920792  1.        ]
```

```
tpr:  [ 0.008  0.02   0.028 ...,  1.     1.     1.   ]
threshold:  [ 0.99999984  0.99999595  0.99999542 ...,  0.0116536   0.01165196
  0.00169226]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
================================================================
Optimized Random Forest hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.97      0.95       505
        Yes       0.94      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 92.85%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.967524752475
fpr:  [ 0.          0.          0.          ...,  0.91287129  0.97029703  1.        ]
tpr:  [ 0.08   0.084  0.092 ...,  0.992  1.     1.   ]
threshold:  [ 1.          0.9978022   0.99615385 ...,  0.00923077  0.00769231  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.96      0.93       505
        Yes       0.91      0.79      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 90.33%
---------------------------
Confusion Matrix
---------------------------
[[485  20]
 [ 53 197]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.877881188119
fpr:  [ 0.          0.04950495  0.05742574  0.05742574  1.        ]
tpr:  [ 0.     0.788  0.8    0.808  1.   ]
threshold:  [ 2.          1.          0.5         0.33333333  0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.44%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 29 221]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.929128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.884  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 113.25453400611877 seconds
=================================================================
```

Combined data – pre-processing with porter stemmer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.96      0.93       505
        Yes       0.90      0.78      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 90.07%
---------------------------
Confusion Matrix
---------------------------
[[484  21]
 [ 54 196]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.955057425743
fpr:  [ 0.          0.0039604   0.00990099  0.04158416  0.1049505   0.20990099
  1.        ]
tpr:  [ 0.     0.588  0.692  0.784  0.896  0.964  1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.82      0.96      0.89       505
        Yes       0.89      0.57      0.69       250

avg / total       0.84      0.83      0.82       755

The accuracy score is 83.31%
---------------------------
Confusion Matrix
---------------------------
[[487  18]
 [108 142]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.927148514851
fpr:  [ 0.          0.          0.         ...,  0.99009901  0.99405941  1.        ]
tpr:  [ 0.008  0.016  0.028 ...,  1.     1.     1.   ]
threshold:  [ 0.99999998  0.99999974  0.99999842 ...,  0.00714477  0.00646969
  0.00311335]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       505
        Yes       0.92      0.80      0.86       250

avg / total       0.91      0.91      0.91       755

The accuracy score is 91.26%
---------------------------
Confusion Matrix
---------------------------
[[488  17]
 [ 49 201]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.89019009901
fpr:  [ 0.          0.03366337  0.04158416  1.        ]
tpr:  [ 0.     0.804  0.816  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.99      0.95       505
        Yes       0.97      0.82      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.11%
---------------------------
Confusion Matrix
---------------------------
[[498   7]
 [ 45 205]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.956328712871
fpr:  [ 0.          0.          0.0039604  ...,  0.41980198  0.42574257  1.        ]
tpr:  [ 0.316  0.324  0.5   ...,  0.96   0.96   1.   ]
threshold:  [ 1.    0.95  0.9  ...,  0.1   0.05  0.  ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.62      0.73       505
        Yes       0.53      0.88      0.66       250

avg / total       0.78      0.70      0.71       755

The accuracy score is 70.20%
---------------------------
Confusion Matrix
---------------------------
[[311 194]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.825548514851
fpr:  [ 0.          0.18217822  0.19009901 ...,  0.95841584  0.95841584  1.        ]
tpr:  [ 0.     0.692  0.704 ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   3.16392206e-089   3.69167559e-092   3.46692607e-166]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.95      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.44%
---------------------------
Confusion Matrix
---------------------------
[[494  11]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.972542574257
fpr:  [ 0.          0.          0.         ...,  0.85544554  0.85940594  1.        ]
tpr:  [ 0.008  0.012  0.02  ...,  1.     1.     1.   ]
threshold:  [ 0.99999978  0.99999886  0.99999791 ...,  0.01258842  0.01249669
  0.0017346 ]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       505
        Yes       0.95      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.25%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.963996039604
fpr:  [ 0.          0.          0.         ...,  0.96633663  0.96831683  1.        ]
tpr:  [ 0.104  0.108  0.144 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99423077  0.99230769 ...,  0.00769231  0.0025641   0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.93      0.92       505
        Yes       0.86      0.81      0.83       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 89.27%
---------------------------
Confusion Matrix
---------------------------
[[471  34]
 [ 47 203]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.872506930693
fpr:  [ 0.          0.04554455  0.05148515  1.        ]
tpr:  [ 0.     0.788  0.792  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.97      0.96       505
        Yes       0.95      0.90      0.92       250

avg / total       0.95      0.95      0.95       755
```

```
The accuracy score is 94.83%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 26 224]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.935128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.896  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 140.83306217193604 seconds
=================================================================
```

Combined data – pre-processing with lemmatization

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.95      0.93       505
        Yes       0.90      0.79      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 89.93%
---------------------------
Confusion Matrix
---------------------------
[[482  23]
 [ 53 197]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950926732673
fpr:  [ 0.          0.0039604   0.01980198  0.04554455  0.11089109  0.20990099
  1.        ]
tpr:  [ 0.     0.592  0.66   0.788  0.896  0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.96      0.88       505
        Yes       0.87      0.52      0.66       250

avg / total       0.83      0.82      0.80       755

The accuracy score is 81.72%
---------------------------
Confusion Matrix
---------------------------
```

```
[[486  19]
 [119 131]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.915271287129
fpr:  [ 0.          0.          0.         ...,  0.99405941  0.9980198   1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.99999872  0.99999026  0.99999012 ...,  0.01239373  0.01214993
  0.01159846]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.96      0.94       505
        Yes       0.91      0.84      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 91.92%
---------------------------
Confusion Matrix
---------------------------
[[485  20]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.899631683168
fpr:  [ 0.          0.03960396  0.04554455  1.        ]
tpr:  [ 0.     0.836  0.84   1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.95       505
        Yes       0.95      0.82      0.88       250

avg / total       0.93      0.93      0.92       755

The accuracy score is 92.58%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 46 204]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.968712871287
fpr:  [ 0.          0.0039604   0.00792079 ...,  0.41584158  0.42178218  1.        ]
tpr:  [ 0.     0.324  0.512 ...,  0.98   0.98   1.   ]
threshold:  [ 2.     1.     0.9    ...,  0.1    0.025  0.   ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support
        No       0.92      0.62      0.74       505
       Yes       0.54      0.90      0.67       250

avg / total       0.80      0.71      0.72       755

The accuracy score is 71.26%
---------------------------
Confusion Matrix
---------------------------
[[314 191]
 [ 26 224]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.845722772277
fpr:  [ 0.          0.18019802  0.18613861 ...,  0.96237624  0.96237624  1.        ]
tpr:  [ 0.     0.756  0.76  ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   5.55132124e-099   2.99145885e-099   2.13113933e-172]
============================================================
Optimized SVC hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

        No       0.94      0.98      0.96       505
       Yes       0.95      0.88      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.57%
---------------------------
Confusion Matrix
---------------------------
[[493  12]
 [ 29 221]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973310891089
fpr:  [ 0.          0.          0.         ...,  0.89306931  0.8970297   1.        ]
tpr:  [ 0.004  0.012  0.016 ...,  1.     1.     1.   ]
threshold:  [ 0.99999932  0.99999785  0.99999465 ...,  0.01259786  0.01235054
  0.00382042]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
============================================================
Optimized Random Forest hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

        No       0.92      0.98      0.95       505
       Yes       0.95      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.11%
---------------------------
Confusion Matrix
```

```
---------------------------
[[494  11]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965742574257
fpr:  [ 0.          0.          0.         ...,  0.95445545  0.95643564  1.        ]
tpr:  [ 0.092  0.144  0.156 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99230769  0.98461538 ...,  0.00769231  0.00480769  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.89      0.95      0.92       505
        Yes       0.89      0.77      0.83       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 89.40%
---------------------------
Confusion Matrix
---------------------------
[[482  23]
 [ 57 193]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.874384158416
fpr:  [ 0.          0.03762376  0.04356436  1.        ]
tpr:  [ 0.     0.78   0.788  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.44%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 29 221]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.929128712871
```

```
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.884  1.    ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 131.33825516700745 seconds
=================================================================
```

Combined data – pre-processing with bi-grams

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.95      0.93       505
        Yes       0.89      0.79      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 89.80%
---------------------------
Confusion Matrix
---------------------------
[[480  25]
 [ 52 198]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950403960396
fpr:  [ 0.          0.00594059  0.02178218  0.04950495  0.11089109  0.21188119
  1.        ]
tpr:  [ 0.     0.596  0.676  0.792  0.9    0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.96      0.87       505
        Yes       0.88      0.50      0.64       250

avg / total       0.82      0.81      0.80       755

The accuracy score is 81.19%
---------------------------
Confusion Matrix
---------------------------
[[487  18]
 [124 126]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.913196039604
fpr:  [ 0.          0.          0.         ...,  0.99009901  0.9980198   1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.9999999   0.99999805  0.99999148 ...,  0.0151756   0.01078191
  0.00998367]
```

```
=====================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       505
        Yes       0.93      0.81      0.86       250

avg / total       0.92      0.92      0.91       755

The accuracy score is 91.52%
---------------------------
Confusion Matrix
---------------------------
[[489  16]
 [ 48 202]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.893398019802
fpr:  [ 0.          0.03168317  0.03762376  1.        ]
tpr:  [ 0.     0.808  0.82   1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=====================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.94       505
        Yes       0.94      0.81      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 92.19%
---------------------------
Confusion Matrix
---------------------------
[[493  12]
 [ 47 203]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.957104950495
fpr:  [ 0.          0.0039604   0.0039604  ...,  0.4039604   0.40990099  1.        ]
tpr:  [ 0.     0.38   0.388 ...,  0.964  0.964  1.   ]
threshold:  [ 2.          1.          0.925      ...,  0.02857143  0.025       0.        ]
=====================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.62      0.75       505
        Yes       0.55      0.92      0.68       250

avg / total       0.81      0.72      0.73       755

The accuracy score is 71.92%
---------------------------
Confusion Matrix
---------------------------
```

```
[[313 192]
 [ 20 230]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.848158415842
fpr:  [ 0.          0.18415842  0.18613861 ...,  0.96039604  0.96039604  1.        ]
tpr:  [ 0.     0.768  0.776 ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   5.18602059e-100   1.21175538e-101   1.30061825e-176]
===================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.98      0.96       505
        Yes       0.95      0.89      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.70%
---------------------------
Confusion Matrix
---------------------------
[[493  12]
 [ 28 222]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973310891089
fpr:  [ 0.          0.          0.         ...,  0.88118812  0.88514851  1.        ]
tpr:  [ 0.008  0.012  0.02  ...,  1.     1.     1.   ]
threshold:  [ 0.99999975  0.99999733  0.99999412 ...,  0.01346437  0.01343971
  0.00299762]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
===================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       505
        Yes       0.96      0.83      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.11%
---------------------------
Confusion Matrix
---------------------------
[[496   9]
 [ 43 207]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.968027722772
fpr:  [ 0.          0.          0.         ...,  0.94851485  0.95247525  1.        ]
tpr:  [ 0.108  0.136  0.184 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99230769  0.97692308 ...,  0.00769231  0.00153846  0.        ]
```

```
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.88      0.95      0.92       505
        Yes       0.89      0.75      0.81       250

avg / total       0.88      0.88      0.88       755

The accuracy score is 88.48%
---------------------------
Confusion Matrix
---------------------------
[[481  24]
 [ 63 187]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.86815049505
fpr:  [ 0.          0.06534653  0.07128713  0.07128713  1.        ]
tpr:  [ 0.     0.788  0.796  0.804  1.   ]
threshold:  [ 2.          1.          0.5         0.33333333  0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.925128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 133.25876235961914 seconds
================================================================
```

Combined data – pre-processing with tri-grams

```
============================================================
KNN classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.95      0.93       505
        Yes       0.89      0.79      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 89.80%
---------------------------
Confusion Matrix
---------------------------
[[480  25]
 [ 52 198]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950403960396
fpr:  [ 0.          0.00594059  0.02178218  0.04950495  0.11089109  0.21188119
  1.        ]
tpr:  [ 0.     0.596  0.676  0.792  0.9    0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
============================================================
SVC classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.96      0.87       505
        Yes       0.88      0.50      0.64       250

avg / total       0.82      0.81      0.80       755

The accuracy score is 81.19%
---------------------------
Confusion Matrix
---------------------------
[[487  18]
 [124 126]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.913196039604
fpr:  [ 0.          0.          0.         ...,  0.99009901  0.9980198   1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.99999986  0.99999728  0.99998848 ...,  0.01640294  0.01174137
  0.01089033]
============================================================
Decision Tree classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.97      0.94       505
        Yes       0.93      0.79      0.86       250

avg / total       0.91      0.91      0.91       755
```

```
The accuracy score is 91.13%
---------------------------
Confusion Matrix
---------------------------
[[490  15]
 [ 52 198]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.886293069307
fpr:  [ 0.          0.02970297  0.02970297  0.03564356  1.        ]
tpr:  [ 0.     0.788  0.792  0.804  1.   ]
threshold:  [ 2.          1.          0.66666667  0.5         0.        ]
==============================================================
Random Forest classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.94       505
        Yes       0.95      0.80      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 92.19%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 49 201]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953287128713
fpr:  [ 0.          0.0039604   0.0039604  ...,  0.41188119  0.41584158  1.        ]
tpr:  [ 0.     0.34   0.344 ...,  0.96   0.96   1.   ]
threshold:  [ 2.     1.     0.925 ...,  0.1    0.05   0.   ]
==============================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.62      0.75       505
        Yes       0.55      0.92      0.68       250

avg / total       0.81      0.72      0.73       755

The accuracy score is 71.92%
---------------------------
Confusion Matrix
---------------------------
[[313 192]
 [ 20 230]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.848158415842
fpr:  [ 0.          0.18415842  0.18613861 ...,  0.96039604  0.96039604  1.        ]
tpr:  [ 0.     0.768  0.776 ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   5.18602059e-100   1.21175538e-101   1.30061825e-176]
```

```
============================================================
Optimized SVC hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.98      0.96       505
        Yes       0.95      0.89      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.70%
--------------------------
Confusion Matrix
--------------------------
[[493  12]
 [ 28 222]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.973310891089
fpr:  [ 0.          0.          0.         ...,  0.88118812  0.88514851  1.        ]
tpr:  [ 0.008  0.012  0.02  ...,  1.     1.     1.   ]
threshold:  [ 0.99999976  0.99999738  0.99999419 ...,  0.0121842   0.01216166
  0.00267297]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
============================================================
Optimized Random Forest hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       505
        Yes       0.95      0.83      0.88       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 92.85%
--------------------------
Confusion Matrix
--------------------------
[[494  11]
 [ 43 207]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.965572277228
fpr:  [ 0.          0.          0.         ...,  0.88910891  0.95247525  1.        ]
tpr:  [ 0.104  0.148  0.152 ...,  0.992  0.996  1.   ]
threshold:  [ 1.          0.99230769  0.98615385 ...,  0.00879121  0.00769231  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
============================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.94      0.92       505
        Yes       0.86      0.78      0.82       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 88.74%
---------------------------
Confusion Matrix
---------------------------
[[474  31]
 [ 54 196]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.866542574257
fpr:  [ 0.          0.05742574  0.06336634  1.        ]
tpr:  [ 0.     0.788  0.792  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
============================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.925128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

============================================================
Program ran in 133.772207736969 seconds
============================================================
```

Combined data – pre-processing with 4-grams

```
============================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.95      0.93       505
        Yes       0.89      0.79      0.84       250
```

```
avg / total       0.90      0.90      0.90       755

The accuracy score is 89.80%
---------------------------
Confusion Matrix
---------------------------
[[480  25]
 [ 52 198]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950403960396
fpr:  [ 0.          0.00594059  0.02178218  0.04950495  0.11089109  0.21188119
  1.        ]
tpr:  [ 0.     0.596  0.676  0.792  0.9    0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
================================================================
SVC classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.80      0.96      0.87       505
        Yes       0.88      0.50      0.64       250

avg / total       0.82      0.81      0.80       755

The accuracy score is 81.19%
---------------------------
Confusion Matrix
---------------------------
[[487  18]
 [124 126]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.913196039604
fpr:  [ 0.          0.          0.         ...,  0.99009901  0.9980198   1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.99999987  0.99999753  0.99998945 ...,  0.01598344  0.01141499
  0.01058219]
================================================================
Decision Tree classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       505
        Yes       0.92      0.80      0.86       250

avg / total       0.91      0.91      0.91       755

The accuracy score is 91.26%
---------------------------
Confusion Matrix
---------------------------
[[488  17]
 [ 49 201]]
---------------------------
ROC Curve
---------------------------
```

```
Area under the curve:  0.888451485149
fpr:  [ 0.          0.03366337  0.03960396  1.        ]
tpr:  [ 0.     0.804  0.812  1.    ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.98      0.94       505
        Yes       0.95      0.80      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 91.92%
---------------------------
Confusion Matrix
---------------------------
[[494  11]
 [ 50 200]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.966491089109
fpr:  [ 0.          0.          0.00792079 ...,  0.38415842  0.38415842  1.        ]
tpr:  [ 0.388  0.392  0.564 ...,  0.976  0.98   1.   ]
threshold:  [ 1.      0.95    0.9    ...,  0.0375  0.025   0.    ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.62      0.75       505
        Yes       0.55      0.92      0.68       250

avg / total       0.81      0.72      0.73       755

The accuracy score is 71.92%
---------------------------
Confusion Matrix
---------------------------
[[313 192]
 [ 20 230]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.848158415842
fpr:  [ 0.          0.18415842  0.18613861 ...,  0.96039604  0.96039604  1.        ]
tpr:  [ 0.     0.768  0.776 ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   5.18602059e-100   1.21175538e-101   1.30061825e-176]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.98      0.96       505
        Yes       0.95      0.89      0.92       250

avg / total       0.95      0.95      0.95       755
```

```
The accuracy score is 94.70%
---------------------------
Confusion Matrix
---------------------------
[[493  12]
 [ 28 222]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973310891089
fpr:  [ 0.          0.          0.         ...,  0.88118812  0.88514851  1.        ]
tpr:  [ 0.008  0.012  0.02  ...,  1.     1.     1.   ]
threshold:  [ 0.9999994   0.99999407  0.99998735 ...,  0.01405778  0.01403304
  0.00331982]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.98      0.95       505
        Yes       0.95      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.25%
---------------------------
Confusion Matrix
---------------------------
[[494  11]
 [ 40 210]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.968372277228
fpr:  [ 0.          0.          0.         ...,  0.9049505   0.95841584  1.        ]
tpr:  [ 0.08   0.148  0.16  ...,  0.996  1.     1.   ]
threshold:  [ 1.          0.99230769  0.98461538 ...,  0.01153846  0.00769231  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
==================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.96      0.94       505
        Yes       0.91      0.81      0.86       250

avg / total       0.91      0.91      0.91       755

The accuracy score is 91.26%
---------------------------
Confusion Matrix
---------------------------
[[486  19]
```

```
 [ 47 203]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.826487128713
fpr:  [ 0.          0.06732673  0.07722772  0.07722772  1.        ]
tpr:  [ 0.     0.712  0.72   0.724  1.   ]
threshold:  [ 2.          1.          0.5         0.33333333  0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.925128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 129.63768672943115 seconds
=================================================================
```

Combined data – pre-processing with stop word removal

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.80      0.86       505
        Yes       0.68      0.88      0.77       250

avg / total       0.85      0.83      0.83       755

The accuracy score is 82.52%
---------------------------
Confusion Matrix
---------------------------
[[403 102]
 [ 30 220]]
```

```
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.940312871287
fpr:  [ 0.          0.01980198  0.08910891  0.2019802   0.25346535  0.33861386
  1.        ]
tpr:  [ 0.     0.776  0.828  0.88   0.944  0.972  1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.79      0.86      0.82       505
        Yes       0.66      0.53      0.59       250

avg / total       0.74      0.75      0.74       755

The accuracy score is 75.23%
---------------------------
Confusion Matrix
---------------------------
[[435  70]
 [117 133]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.872269306931
fpr:  [ 0.          0.          0.         ...,  0.84158416  0.84158416  1.        ]
tpr:  [ 0.008  0.02   0.024 ...,  0.996  1.     1.   ]
threshold:  [ 0.94197069  0.93173663  0.90378361 ...,  0.10151136  0.10133307
  0.06012833]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.97      0.95       505
        Yes       0.94      0.83      0.88       250

avg / total       0.93      0.93      0.92       755

The accuracy score is 92.58%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 43 207]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.902566336634
fpr:  [ 0.          0.02574257  0.03168317  1.        ]
tpr:  [ 0.     0.828  0.832  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.30%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 30 220]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965021782178
fpr:  [ 0.          0.0039604   0.00792079 ...,  0.2970297   0.3009901   1.        ]
tpr:  [ 0.     0.392  0.68  ...,  0.964  0.964  1.   ]
threshold:  [ 2.    1.    0.9  ...,  0.1   0.05  0.  ]
==============================================================
Bernoulli Naive Bayes classifier results
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.65      0.78       505
        Yes       0.58      0.98      0.73       250

avg / total       0.85      0.76      0.77       755

The accuracy score is 75.89%
---------------------------
Confusion Matrix
---------------------------
[[329 176]
 [  6 244]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.957322772277
fpr:  [ 0.          0.00792079  0.00990099 ...,  0.89306931  0.89306931  1.        ]
tpr:  [ 0.     0.3    0.3   ...,  0.996  1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   8.72099115e-036   6.41327911e-036   6.01342202e-105]
==============================================================
Optimized SVC hyper parameters
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.97      0.96       505
        Yes       0.94      0.89      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.57%
---------------------------
Confusion Matrix
---------------------------
[[491  14]
 [ 27 223]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.96535049505
fpr:  [ 0.          0.          0.         ...,  0.78217822  0.79009901  1.        ]
tpr:  [ 0.004  0.012  0.016 ...,  1.     1.     1.   ]
threshold:  [ 0.99998952  0.99998459  0.99670665 ...,  0.01545642  0.01545097
  0.00796247]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
==================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.97      0.96       505
        Yes       0.94      0.90      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.97%
---------------------------
Confusion Matrix
---------------------------
[[491  14]
 [ 24 226]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.970237623762
fpr:  [ 0.          0.          0.         ...,  0.93861386  0.94059406  1.        ]
tpr:  [ 0.1    0.112  0.136 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99384615  0.99230769 ...,  0.00769231  0.0025641   0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
==================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.94      0.94       505
        Yes       0.87      0.89      0.88       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 92.19%
---------------------------
Confusion Matrix
---------------------------
[[473  32]
 [ 27 223]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.933904950495
fpr:  [ 0.          0.05148515  0.05742574  1.        ]
tpr:  [ 0.     0.916  0.92   1.   ]
threshold:  [ 2.   1.   0.5  0. ]
```

```
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.96      0.97      0.96       505
        Yes       0.94      0.91      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 95.10%
---------------------------
Confusion Matrix
---------------------------
[[490  15]
 [ 22 228]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.941148514851
fpr:  [ 0.          0.02970297  1.        ]
tpr:  [ 0.     0.912  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

================================================================
Program ran in 67.85607624053955 seconds
================================================================
```

Combined data – pre-processing with TF-IDF

```
================================================================
KNN classifier results
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.92      0.91       505
        Yes       0.84      0.81      0.82       250

avg / total       0.88      0.88      0.88       755

The accuracy score is 88.48%
---------------------------
Confusion Matrix
---------------------------
[[465  40]
 [ 47 203]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953398019802
fpr:  [ 0.          0.00792079  0.02574257  0.07920792  0.12871287  0.20594059
  1.        ]
tpr:  [ 0.     0.684  0.744  0.812  0.908  0.968  1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
```

```
=================================================================
SVC classifier results
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.70      1.00      0.82       505
        Yes       1.00      0.13      0.23       250

avg / total       0.80      0.71      0.63       755

The accuracy score is 71.13%
--------------------------
Confusion Matrix
--------------------------
[[505   0]
 [218  32]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.965223762376
fpr:  [ 0.          0.          0.         ...,  0.98613861  0.99009901  1.        ]
tpr:  [ 0.056  0.068  0.092 ...,  1.     1.     1.   ]
threshold:  [ 1.          1.          1.         ...,  0.00856891  0.00829333
  0.00383955]
=================================================================
Decision Tree classifier results
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.97      0.95       505
        Yes       0.93      0.84      0.88       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 92.72%
--------------------------
Confusion Matrix
--------------------------
[[490  15]
 [ 40 210]]
--------------------------
ROC Curve
--------------------------
Area under the curve:  0.908554455446
fpr:  [ 0.          0.02970297  0.03564356  1.        ]
tpr:  [ 0.     0.84   0.848  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.98      0.95       505
        Yes       0.95      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.25%
--------------------------
Confusion Matrix
```

```
---------------------------
[[494  11]
 [ 40 210]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.963065346535
fpr:  [ 0.          0.0039604   0.0039604  ...,  0.36633663  0.37425743  1.        ]
tpr:  [ 0.     0.4    0.404 ...,  0.972  0.972  1.   ]
threshold:  [ 2.          1.          0.95       ...,  0.1         0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.59      0.74       505
        Yes       0.54      0.98      0.70       250

avg / total       0.84      0.72      0.72       755

The accuracy score is 71.92%
---------------------------
Confusion Matrix
---------------------------
[[297 208]
 [  4 246]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.881837623762
fpr:  [ 0.          0.21386139  0.22178218 ...,  0.68910891  0.97425743  1.        ]
tpr:  [ 0.     0.96   0.964 ...,  1.     1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   3.70177066e-074   5.24758958e-311   0.00000000e+000]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.95      0.88      0.91       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.57%
---------------------------
Confusion Matrix
---------------------------
[[494  11]
 [ 30 220]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.963259405941
fpr:  [ 0.          0.          0.         ...,  0.78415842  0.78415842  1.        ]
tpr:  [ 0.008  0.024  0.032 ...,  0.996  1.     1.   ]
threshold:  [ 0.99999374  0.99620278  0.99598854 ...,  0.01875256  0.01875138
  0.01029408]
```

```
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=============================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.970091089109
fpr:  [ 0.          0.          0.         ...,  0.87920792  0.96237624  1.        ]
tpr:  [ 0.1    0.108  0.132 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99538462  0.99230769 ...,  0.00923077  0.00769231  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
=============================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.97      0.94       505
        Yes       0.92      0.80      0.86       250

avg / total       0.91      0.91      0.91       755

The accuracy score is 91.26%
---------------------------
Confusion Matrix
---------------------------
[[488  17]
 [ 49 201]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.900697029703
fpr:  [ 0.          0.04950495  0.05544554  1.        ]
tpr:  [ 0.     0.848  0.852  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=============================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support
         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.90       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 93.91%
---------------------------
Confusion Matrix
---------------------------
[[490  15]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.923148514851
fpr:  [ 0.          0.02970297  1.        ]
tpr:  [ 0.     0.876  1.    ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 152.67092442512512 seconds
=================================================================
```

Combined data – pre-processing with TF-IDF + porter stemmer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.90      0.94      0.92       505
        Yes       0.86      0.80      0.83       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 89.01%
---------------------------
Confusion Matrix
---------------------------
[[473  32]
 [ 51 199]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.952499009901
fpr:  [ 0.          0.0039604   0.01386139  0.06336634  0.11683168  0.23762376
  1.        ]
tpr:  [ 0.     0.664  0.728  0.796  0.9    0.964  1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.70      1.00      0.82       505
        Yes       1.00      0.13      0.23       250
```

```
avg / total       0.80      0.71      0.63       755

The accuracy score is 71.26%
---------------------------
Confusion Matrix
---------------------------
[[505   0]
 [217  33]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.963861386139
fpr:  [ 0.          0.          0.         ...,  0.96435644  0.9960396   1.        ]
tpr:  [ 0.08   0.104  0.116 ...,  1.     1.     1.   ]
threshold:  [ 1.          1.          1.         ...,  0.01156662  0.00462716
  0.00387667]
=====================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.98      0.95       505
        Yes       0.96      0.85      0.90       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 93.77%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 37 213]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.91761980198
fpr:  [ 0.          0.01980198  0.02574257  1.        ]
tpr:  [ 0.     0.852  0.856  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=====================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.87      0.90       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 93.91%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 33 217]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973192079208
fpr:  [ 0.          0.00792079  0.00990099 ...,  0.4         0.4039604   1.        ]
```

```
tpr:  [ 0.388  0.576  0.7   ...,  0.984  0.984  1.   ]
threshold:  [ 1.         0.9        0.8        ...,  0.03636364  0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.59      0.74       505
        Yes       0.54      0.98      0.70       250

avg / total       0.84      0.72      0.72       755

The accuracy score is 71.92%
---------------------------
Confusion Matrix
---------------------------
[[298 207]
 [  5 245]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.860605940594
fpr:  [ 0.          0.25742574  0.27524752 ...,  0.6970297   0.97425743  1.        ]
tpr:  [ 0.     0.964  0.964 ...,  1.     1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   1.58716559e-079   3.38676095e-310   0.00000000e+000]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.95      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.44%
---------------------------
Confusion Matrix
---------------------------
[[494  11]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.965786138614
fpr:  [ 0.         0.         0.         ...,  0.7960396  0.8        1.       ]
tpr:  [ 0.008  0.024  0.032 ...,  1.     1.     1.   ]
threshold:  [ 0.99999715  0.99716053  0.99694021 ...,  0.01529644  0.01529522
  0.00802569]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.97      0.96       505
        Yes       0.95      0.90      0.92       250
```

```
avg / total       0.95      0.95      0.95       755

The accuracy score is 94.83%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 26 224]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.972974257426
fpr:  [ 0.          0.          0.         ...,  0.94851485  0.95247525  1.        ]
tpr:  [ 0.108  0.116  0.12  ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.9974359   0.99487179 ...,  0.00769231  0.00153846  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
================================================================
Optimized Decision Tree hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.95      0.94       505
        Yes       0.90      0.84      0.87       250

avg / total       0.91      0.92      0.91       755

The accuracy score is 91.52%
---------------------------
Confusion Matrix
---------------------------
[[482  23]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.907497029703
fpr:  [ 0.          0.03960396  0.04554455  1.        ]
tpr:  [ 0.     0.844  0.856  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
================================================================
Optimized KNN hyper parameters
----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.95       505
        Yes       0.93      0.88      0.90       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 93.77%
---------------------------
Confusion Matrix
---------------------------
```

```
[[489  16]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.922158415842
fpr:  [ 0.          0.03168317  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 187.1040117740631 seconds
=================================================================
```

Combined data – pre-processing with TF-IDF + lemmatizer

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.92      0.92       505
        Yes       0.83      0.82      0.83       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 88.74%
---------------------------
Confusion Matrix
---------------------------
[[464  41]
 [ 44 206]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.950194059406
fpr:  [ 0.          0.00792079  0.02574257  0.08118812  0.12673267  0.21782178
  1.        ]
tpr:  [ 0.     0.66   0.744  0.824  0.916  0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.70      1.00      0.82       505
        Yes       1.00      0.13      0.23       250

avg / total       0.80      0.71      0.63       755

The accuracy score is 71.26%
---------------------------
Confusion Matrix
---------------------------
[[505   0]
 [217  33]]
---------------------------
ROC Curve
```

```
---------------------------
Area under the curve:  0.96276039604
fpr:  [ 0.          0.          0.         ...,  0.99207921  0.9960396   1.        ]
tpr:  [ 0.056  0.068  0.084 ...,  1.     1.     1.   ]
threshold:  [ 1.          1.          1.         ...,  0.00681916  0.00646311
  0.00303315]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.97      0.94       505
        Yes       0.93      0.82      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 92.05%
---------------------------
Confusion Matrix
---------------------------
[[489  16]
 [ 44 206]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.901421782178
fpr:  [ 0.          0.03168317  0.03762376  0.03762376  1.        ]
tpr:  [ 0.     0.824  0.828  0.836  1.   ]
threshold:  [ 2.          1.          0.5         0.33333333  0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.96      0.94       505
        Yes       0.92      0.82      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 91.79%
---------------------------
Confusion Matrix
---------------------------
[[487  18]
 [ 44 206]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.955992079208
fpr:  [ 0.          0.          0.01188119 ...,  0.43762376  0.44158416  1.        ]
tpr:  [ 0.372  0.376  0.572 ...,  0.968  0.968  1.   ]
threshold:  [ 1.     0.94   0.9   ...,  0.04   0.025  0.   ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.59      0.74       505
        Yes       0.54      0.98      0.70       250
```

```
avg / total       0.84      0.72      0.73       755

The accuracy score is 72.05%
---------------------------
Confusion Matrix
---------------------------
[[299 206]
 [  5 245]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.878526732673
fpr:  [ 0.          0.21980198  0.24158416 ...,  0.68910891  0.97425743  1.        ]
tpr:  [ 0.     0.96   0.964 ...,  1.     1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   4.29184198e-074   2.32540859e-310   0.00000000e+000]
===============================================================
Optimized SVC hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.95      0.87      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[493  12]
 [ 32 218]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.966776237624
fpr:  [ 0.          0.          0.         ...,  0.77227723  0.77623762  1.        ]
tpr:  [ 0.008  0.02   0.028 ...,  1.     1.     1.   ]
threshold:  [ 0.99999614  0.99694328  0.99655393 ...,  0.01756682  0.01756579
  0.00880059]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
===============================================================
Optimized Random Forest hyper parameters
---------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.97      0.96       505
        Yes       0.94      0.90      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.70%
---------------------------
Confusion Matrix
---------------------------
[[491  14]
 [ 26 224]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.970372277228
fpr:  [ 0.          0.          0.         ...,  0.95049505  0.95247525  1.        ]
tpr:  [ 0.124  0.152  0.2   ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99230769  0.98461538 ...,  0.00769231  0.0025641   0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.91      0.93      0.92       505
        Yes       0.86      0.82      0.84       250

avg / total       0.90      0.90      0.90       755

The accuracy score is 89.80%
---------------------------
Confusion Matrix
---------------------------
[[472  33]
 [ 44 206]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.901667326733
fpr:  [ 0.          0.04356436  0.04950495  1.        ]
tpr:  [ 0.     0.844  0.848  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
=================================================================
Optimized KNN hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.17%
---------------------------
Confusion Matrix
---------------------------
[[492  13]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.925128712871
fpr:  [ 0.          0.02574257  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
```

```
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

=================================================================
Program ran in 168.949116230011 seconds
=================================================================
```

Combined data – pre-processing with TF-IDF + tri-grams

```
=================================================================
KNN classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.92      0.92       505
        Yes       0.84      0.83      0.84       250

avg / total       0.89      0.89      0.89       755

The accuracy score is 89.14%
---------------------------
Confusion Matrix
---------------------------
[[465  40]
 [ 42 208]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.948534653465
fpr:  [ 0.          0.01188119  0.02772277  0.07920792  0.14257426  0.24356436
  1.        ]
tpr:  [ 0.     0.692  0.756  0.832  0.924  0.96   1.   ]
threshold:  [ 2.   1.   0.8  0.6  0.4  0.2  0. ]
=================================================================
SVC classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.68      1.00      0.81       505
        Yes       1.00      0.03      0.06       250

avg / total       0.78      0.68      0.56       755

The accuracy score is 67.95%
---------------------------
Confusion Matrix
---------------------------
[[505   0]
 [242   8]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.953817821782
fpr:  [ 0.          0.          0.         ...,  0.98415842  0.99207921  1.        ]
tpr:  [ 0.032  0.036  0.044 ...,  1.     1.     1.   ]
threshold:  [ 1.          1.          1.         ...,  0.01316644  0.01162215
  0.00566082]
=================================================================
Decision Tree classifier results
-----------------------------------------------------------------
```

```
             precision    recall  f1-score   support

         No       0.92      0.97      0.94       505
        Yes       0.93      0.82      0.87       250

avg / total       0.92      0.92      0.92       755

The accuracy score is 92.19%
---------------------------
Confusion Matrix
---------------------------
[[490  15]
 [ 44 206]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.898566336634
fpr:  [ 0.          0.02970297  0.03564356  1.        ]
tpr:  [ 0.     0.824  0.828  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.98      0.95       505
        Yes       0.95      0.84      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 93.25%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 41 209]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.964617821782
fpr:  [ 0.          0.0039604   0.0039604  ...,  0.42772277  0.43168317  1.        ]
tpr:  [ 0.404  0.588  0.592 ...,  0.968  0.968  1.   ]
threshold:  [ 1.          0.9         0.86666667 ...,  0.1         0.03333333  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.99      0.59      0.74       505
        Yes       0.54      0.98      0.70       250

avg / total       0.84      0.72      0.73       755

The accuracy score is 72.19%
---------------------------
Confusion Matrix
---------------------------
[[299 206]
 [  4 246]]
---------------------------
```

```
ROC Curve
---------------------------
Area under the curve:  0.828708910891
fpr:  [ 0.          0.32475248  0.32871287 ...,  0.72079208  0.94653465  1.        ]
tpr:  [ 0.     0.968  0.968 ...,  1.     1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   5.45418557e-131   1.82667575e-306   0.00000000e+000]
=================================================================
Optimized SVC hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.98      0.96       505
        Yes       0.96      0.88      0.92       250

avg / total       0.95      0.95      0.95       755

The accuracy score is 94.83%
---------------------------
Confusion Matrix
---------------------------
[[495  10]
 [ 29 221]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.969588118812
fpr:  [ 0.          0.          0.         ...,  0.75643564  0.76039604  1.        ]
tpr:  [ 0.008  0.016  0.024 ...,  1.     1.     1.   ]
threshold:  [ 0.99999863  0.99999444  0.99998695 ...,  0.01617051  0.01617023
  0.00728475]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.96       505
        Yes       0.94      0.88      0.91       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 94.04%
---------------------------
Confusion Matrix
---------------------------
[[491  14]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.973192079208
fpr:  [ 0.          0.          0.         ...,  0.92475248  0.96633663  1.        ]
tpr:  [ 0.096  0.208  0.212 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.98461538  0.98012821 ...,  0.00923077  0.00769231  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
============================================================
Optimized Decision Tree hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.95      0.95       505
        Yes       0.90      0.87      0.89       250

avg / total       0.93      0.93      0.93       755

The accuracy score is 92.72%
---------------------------
Confusion Matrix
---------------------------
[[482  23]
 [ 32 218]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.898661386139
fpr:  [ 0.          0.04554455  0.05148515  1.        ]
tpr:  [ 0.     0.84   0.844  1.   ]
threshold:  [ 2.   1.   0.5  0. ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
============================================================
Optimized KNN hyper parameters
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.97      0.95       505
        Yes       0.93      0.88      0.90       250

avg / total       0.94      0.94      0.94       755

The accuracy score is 93.77%
---------------------------
Confusion Matrix
---------------------------
[[489  16]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.922158415842
fpr:  [ 0.          0.03168317  1.        ]
tpr:  [ 0.     0.876  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

============================================================
Program ran in 219.0989751815796 seconds
============================================================
```

Combined data – pre-processing with TF-IDF + stop word removal

```
============================================================
KNN classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.84      0.88       505
        Yes       0.73      0.86      0.79       250

avg / total       0.86      0.85      0.85       755

The accuracy score is 84.64%
-------------------------
Confusion Matrix
-------------------------
[[425  80]
 [ 36 214]]
-------------------------
ROC Curve
-------------------------
Area under the curve:  0.924962376238
fpr:  [ 0.          0.00792079  0.15841584  0.22376238  0.28712871  1.        ]
tpr:  [ 0.276  0.484  0.856  0.924  0.972  1.   ]
threshold:  [ 1.   0.8  0.6  0.4  0.2  0. ]
============================================================
SVC classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.72      0.89      0.80       505
        Yes       0.58      0.31      0.40       250

avg / total       0.67      0.70      0.67       755

The accuracy score is 69.67%
-------------------------
Confusion Matrix
-------------------------
[[449  56]
 [173  77]]
-------------------------
ROC Curve
-------------------------
Area under the curve:  0.857924752475
fpr:  [ 0.          0.10891089  0.11089109 ...,  0.99207921  0.9960396   1.        ]
tpr:  [ 0.     0.212  0.212 ...,  1.     1.     1.   ]
threshold:  [ 1.81809613  0.81809613  0.72748636 ...,  0.15126235  0.15096418
  0.12641415]
============================================================
Decision Tree classifier results
------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.95      0.86      0.90       505
        Yes       0.77      0.90      0.83       250

avg / total       0.89      0.88      0.88       755

The accuracy score is 87.55%
-------------------------
```

```
Confusion Matrix
---------------------------
[[436  69]
 [ 25 225]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.92676039604
fpr:  [ 0.          0.00990099  0.01188119 ...,  0.15841584  0.15841584  1.        ]
tpr:  [ 0.     0.628  0.632 ...,  0.916  0.92   1.   ]
threshold:  [ 2.          1.          0.71428571 ...,  0.30769231  0.25        0.        ]
=================================================================
Random Forest classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.86      0.89       505
        Yes       0.75      0.86      0.80       250

avg / total       0.87      0.86      0.86       755

The accuracy score is 85.83%
---------------------------
Confusion Matrix
---------------------------
[[434  71]
 [ 36 214]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.937976237624
fpr:  [ 0.          0.          0.0019802  ...,  0.37425743  0.38019802  1.        ]
tpr:  [ 0.272  0.276  0.424 ...,  0.976  0.98   1.   ]
threshold:  [ 1.          0.95388128  0.9        ...,  0.04        0.02631579  0.        ]
=================================================================
Bernoulli Naive Bayes classifier results
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.98      0.60      0.75       505
        Yes       0.55      0.98      0.71       250

avg / total       0.84      0.73      0.73       755

The accuracy score is 72.85%
---------------------------
Confusion Matrix
---------------------------
[[305 200]
 [  5 245]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.919267326733
fpr:  [ 0.          0.01584158  0.01782178 ...,  0.76435644  0.76831683  1.        ]
tpr:  [ 0.     0.376  0.376 ...,  1.     1.     1.   ]
threshold:  [  2.00000000e+000   1.00000000e+000   1.00000000e+000 ...,
   1.37660375e-042   8.20521539e-043   2.59089248e-205]
=================================================================
Optimized SVC hyper parameters
```

```
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.94      0.85      0.89       505
        Yes       0.75      0.89      0.81       250

avg / total       0.88      0.86      0.87       755

The accuracy score is 86.36%
---------------------------
Confusion Matrix
---------------------------
[[430  75]
 [ 28 222]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.87243960396
fpr:  [ 0.          0.          0.         ...,  0.99207921  0.99207921  1.        ]
tpr:  [ 0.004  0.08   0.088 ...,  0.996  1.     1.   ]
threshold:  [ 0.93850269  0.81151542  0.81147303 ...,  0.0440058   0.04115181
  0.03223761]
Parameters were:  {'C': 10, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': None,
'degree': 3, 'gamma': 1, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': None, 'shrinking': True,
'tol': 0.001, 'verbose': False}
=================================================================
Optimized Random Forest hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.86      0.89       505
        Yes       0.75      0.88      0.81       250

avg / total       0.87      0.86      0.87       755

The accuracy score is 86.36%
---------------------------
Confusion Matrix
---------------------------
[[433  72]
 [ 31 219]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.944297029703
fpr:  [ 0.          0.          0.         ...,  0.75049505  0.75247525  1.        ]
tpr:  [ 0.06   0.064  0.112 ...,  0.996  0.996  1.   ]
threshold:  [ 1.          0.99661734  0.99230769 ...,  0.0025641   0.00153846  0.        ]
Parameters were:  {'bootstrap': True, 'class_weight': None, 'criterion': 'entropy', 'max_depth': None,
'max_features': 'log2', 'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 130, 'n_jobs': 1, 'oob_score': False,
'random_state': None, 'verbose': 0, 'warm_start': False}
=================================================================
Optimized Decision Tree hyper parameters
-----------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.93      0.85      0.89       505
        Yes       0.75      0.88      0.81       250
```

```
avg / total       0.87      0.86      0.86       755

The accuracy score is 86.23%
---------------------------
Confusion Matrix
---------------------------
[[431  74]
 [ 30 220]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.902201980198
fpr:  [ 0.          0.01584158  0.01782178 ...,  0.16435644  0.16435644  1.        ]
tpr:  [ 0.     0.572  0.576 ...,  0.876  0.888  1.   ]
threshold:  [ 2.          1.          0.71428571 ...,  0.30769231  0.25        0.        ]
Parameters were:  {'class_weight': None, 'criterion': 'entropy', 'max_depth': None, 'max_features': 5,
'max_leaf_nodes': None, 'min_impurity_split': 1e-07, 'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
==================================================================
Optimized KNN hyper parameters
--------------------------------------------------------------
             precision    recall  f1-score   support

         No       0.92      0.87      0.89       505
        Yes       0.76      0.85      0.80       250

avg / total       0.87      0.86      0.86       755

The accuracy score is 86.09%
---------------------------
Confusion Matrix
---------------------------
[[437  68]
 [ 37 213]]
---------------------------
ROC Curve
---------------------------
Area under the curve:  0.858673267327
fpr:  [ 0.          0.13465347  1.        ]
tpr:  [ 0.     0.852  1.   ]
threshold:  [ 2.  1.  0.]
Parameters were:  {'algorithm': 'auto', 'leaf_size': 30, 'metric': 'euclidean', 'metric_params': None, 'n_jobs': 1,
'n_neighbors': 1, 'p': 1, 'weights': 'uniform'}

==================================================================
Program ran in 73.24604058265686 seconds
==================================================================
```