Final Project Attempt

1. Initial – Kfold with n=5
   1. Avg. Recall: 0.571
   2. Avg. Precision: 0.889
   3. Avg. F-Beta: 0.696
   4. Avg. ROC AUC: 0.955
   5. Time to run: 0:04:56.208785
2. Data cleanup
   1. Remove punctuation, all lowercase
      1. Avg. Recall: 0.566
      2. Avg. Precision: 0.891
      3. Avg. F-Beta: 0.692
      4. Avg. ROC AUC: 0.954
      5. Time to run: 0:03:37.461214
   2. Convert punctuation to spaces, all lowercase
      1. Avg. Recall: 0.571
      2. Avg. Precision: 0.889
      3. Avg. F-Beta: 0.695
      4. Avg. ROC AUC: 0.955
      5. Time to run: 0:03:49.056966
   3. Stop\_words=’english’ with punctuation as spaces - <https://scikit-learn.org/stable/modules/feature_extraction.html#stop-words>
      1. Avg. Recall: 0.559
      2. Avg. Precision: 0.894
      3. Avg. F-Beta: 0.688
      4. Avg. ROC AUC: 0.949
      5. Time to run: 0:03:38.880159
3. Feature Extraction
   1. Words
      1. Unigrams
         1. Avg. Recall: 0.581
         2. Avg. Precision: 0.891
         3. Avg. F-Beta: 0.703
         4. Avg. ROC AUC: 0.958
         5. Time to run: 0:02:35.807145
      2. Unigrams and Bigrams
         1. Avg. Recall: 0.571
         2. Avg. Precision: 0.889
         3. Avg. F-Beta: 0.695
         4. Avg. ROC AUC: 0.955
         5. Time to run: 0:03:47.269405
      3. Unigrams, Bigrams, and Trigrams
         1. Avg. Recall: 0.567
         2. Avg. Precision: 0.886
         3. Avg. F-Beta: 0.691
         4. Avg. ROC AUC: 0.954
         5. Time to run: 0:06:15.367744
   2. Characters
      1. Word boundary bigrams
         1. Avg. Recall: 0.459
         2. Avg. Precision: 0.816
         3. Avg. F-Beta: 0.587
         4. Avg. ROC AUC: 0.909
         5. Time to run: 0:05:16.578190
      2. Word boundary trigrams
         1. Avg. Recall: 0.588
         2. Avg. Precision: 0.886
         3. Avg. F-Beta: 0.707
         4. Avg. ROC AUC: 0.956
         5. Time to run: 0:05:44.735390
      3. Word boundary 4-grams
         1. Avg. Recall: 0.611
         2. Avg. Precision: 0.890
         3. Avg. F-Beta: 0.725
         4. Avg. ROC AUC: 0.963
         5. Time to run: 0:04:51.814777
      4. Word boundary 5-grams
         1. Avg. Recall: 0.610
         2. Avg. Precision: 0.891
         3. Avg. F-Beta: 0.724
         4. Avg. ROC AUC: 0.962
         5. Time to run: 0:04:23.947552
   3. Combined
      1. Word unigrams and Character word boundary 4-grams
         1. Avg. Recall: 0.612
         2. Avg. Precision: 0.893
         3. Avg. F-Beta: 0.726
         4. Avg. ROC AUC: 0.964
         5. Time to run: 0:05:21.840541
      2. Word unigrams and Character word boundary 4-grams with 20k features each
         1. Avg. Recall: 0.602
         2. Avg. Precision: 0.895
         3. Avg. F-Beta: 0.720
         4. Avg. ROC AUC: 0.964
         5. Time to run: 0:06:00.353374
      3. Word unigrams and Character word boundary 4-grams with text length feature
         1. Avg. Recall: 0.415
         2. Avg. Precision: 0.933
         3. Avg. F-Beta: 0.574
         4. Avg. ROC AUC: 0.951
         5. Time to run: 0:05:19.773333
      4. Word unigrams and Character word boundary 5-grams
         1. Avg. Recall: 0.608
         2. Avg. Precision: 0.893
         3. Avg. F-Beta: 0.723
         4. Avg. ROC AUC: 0.963
         5. Time to run: 0:05:03.119479
   4. After with logged\_in added
      1. Avg. Recall: 0.621
      2. Avg. Precision: 0.899
      3. Avg. F-Beta: 0.734
      4. Avg. ROC AUC: 0.964
      5. Time to run: 0:05:56.381002
4. Modeling
   1. LogisticRegression
      1. Avg. Recall: 0.612
      2. Avg. Precision: 0.893
      3. Avg. F-Beta: 0.726
      4. Avg. ROC AUC: 0.964
      5. Time to run: 0:05:21.840541
   2. Multinomial NB
      1. Avg. Recall: 0.602
      2. Avg. Precision: 0.842
      3. Avg. F-Beta: 0.702
      4. Avg. ROC AUC: 0.932
      5. Time to run: 0:04:51.760757
   3. RandomForest
      1. Avg. Recall: 0.522
      2. Avg. Precision: 0.888
      3. Avg. F-Beta: 0.658
      4. Avg. ROC AUC: 0.914
      5. Time to run: 0:08:07.777086
   4. SGDClassifier
      1. Avg. Recall: 0.452
      2. Avg. Precision: 0.946
      3. Avg. F-Beta: 0.611
      4. Avg. ROC AUC: 0.952
      5. Time to run: 0:09:24.646272
5. Parameter Tuning
   1. Parameters
      1. parameters = {  
          'clf\_\_alpha': [.00000001, .000001, .0001, .01, 1, 100],  
          'clf\_\_loss': ['log', 'modified\_huber'],  
          'clf\_\_penalty': ['l2', 'l1', 'elasticnet'],  
          'clf\_\_max\_iter': [1000, 2000],  
          'clf\_\_n\_iter\_no\_change': [5, 10],  
          'clf\_\_class\_weight': ['balanced', None],  
         }
   2. Best parameters
      1. Best Score: 0.945
      2. Best parameters set:
         1. clf\_\_alpha: 0.0001
         2. clf\_\_class\_weight: None
         3. clf\_\_loss: 'modified\_huber'
         4. clf\_\_max\_iter: 1000
         5. clf\_\_n\_iter\_no\_change: 10
         6. clf\_\_penalty: 'elasticnet'
   3. Results
      1. Avg. Recall: 0.613
      2. Avg. Precision: 0.900
      3. Avg. F-Beta: 0.729
      4. Avg. ROC AUC: 0.965
      5. Time to run: 0:05:41.022994