CSI4107 Assignment 2 Report

Part 1

In the first part of the experiment, we used Python to read the twitter messages from the text file. Using Scikit-learn, a machine learning toolkit for Python, we are able to create a $n*m$ matrix for n documents with m features of words using a `CountVectorizer` object.

The `CountVectorizer` object takes an array of text objects representing documents and creates an appropriate matrix representing the counts of token words for each document. Documents are first preprocessed with a preprocess object, and then tokenized with a tokenizer object. Together, these form an analyzer that is called to process every document. We decided to extend the basic analyzer by stemming all the words produced by the preprocessor and tokenizer using the `EnglishStemmer` provided by Natural Language Toolkit (NLTK).

Using the matrix created from this preprocessing, tokenization, and stemming, we were then able to produce a sparse arff file for use in Weka. In the sparse arff file, a twitter document is represented by the index of the token in the bag of words list and the count of that token in that document. Tokens are only specified if they are present in the document. This reduces arff file size as features (i.e. words) not present are not included and it is implied that they are 0 for a given document.

With this arff file, the first run in Weka resulted in the following results from a 10-fold cross validation with the three different classifiers:

Decision Tree:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3455               47.7936 %
Incorrectly Classified Instances      3774               52.2064 %
Kappa statistic                          0.2297
Mean absolute error                      0.28
Root mean squared error                  0.4545
Relative absolute error                 80.8692 %
Root relative squared error            109.2265 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.692     0.365      0.612     0.692     0.649      0.704    positiv
e
                 0.354     0.127      0.379     0.354     0.366      0.636    negativ
e
                 0.224     0.155      0.282     0.224     0.249      0.556    neutral
                 0.344     0.115      0.351     0.344     0.348      0.641    objecti
ve
Weighted Avg.    0.478     0.239      0.46      0.478     0.467      0.651


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2271  363  387  263 |    a = positive
  486  458  214  135 |    b = negative
  634  263  346  304 |    c = neutral
  319  125  281  380 |    d = objective
```

Naive Bayes:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3368               46.5901 %
Incorrectly Classified Instances      3861               53.4099 %
Kappa statistic                          0.244
Mean absolute error                      0.2824
Root mean squared error                  0.445
Relative absolute error                 81.5583 %
Root relative squared error            106.9465 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.582     0.279      0.635     0.582     0.607      0.705    positiv
e
                 0.452     0.183      0.35      0.452     0.395      0.698    negativ
e
                 0.219     0.13       0.315     0.219     0.258      0.597    neutral
                 0.482     0.153      0.363     0.482     0.414      0.745    objecti
ve
Weighted Avg.    0.466     0.211      0.474     0.466     0.465      0.687


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1911  596  363  414 |    a = positive
  369  585  184  155 |    b = negative
  480  361  339  367 |    c = neutral
  251  130  191  533 |    d = objective
```

Support Vector Machine (SMO):

```
=== Stratified cross-validation ===

Correctly Classified Instances        3698               51.1551 %
Incorrectly Classified Instances      3531               48.8449 %
Kappa statistic                          0.2741
Mean absolute error                      0.3202
Root mean squared error                  0.4063
Relative absolute error                 92.476  %
Root relative squared error             97.6539 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.725     0.36       0.626     0.725     0.672      0.716    positiv
e
                 0.371     0.103      0.439     0.371     0.402      0.718    negativ
e
                 0.299     0.162      0.334     0.299     0.316      0.579    neutral
                 0.338     0.094      0.394     0.338     0.364      0.718    objecti
ve
Weighted Avg.    0.512     0.231      0.495     0.512     0.5        0.688



=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2382  291  396  215 |    a = positive
  462  480  253   98 |    b = negative
  601  223  463  260 |    c = neutral
  359  100  273  373 |    d = objective
```

Clearly, the SVM classifier produced the best results with 51.15% correctly classified instances and a precision of 49.5%.

Part 2

When adding features to the bag of words feature set, we first began by counting the amount of smiley-based emoticons and sad-based emoticons. The analysis was carried out on each document using the following code:

```
additional_features["smilies"] = twitter_document.msg_text.count("(:") + twitter_document.msg_text.count(":)") + twitter_document.msg_text.count(":-)") + twitter_document.msg_text.count(":o)") + twitter_document.msg_text.count(":]") + twitter_document.msg_text.count(":3") + twitter_document.msg_text.count(":c)") + 2*twitter_document.msg_text.count(":D") + 2*twitter_document.msg_text.count("C:")
additional_features["exclamations"] = twitter_document.msg_text.count("!")
additional_features["questions"] = twitter_document.msg_text.count("?")
additional_features["sadfaces"] = twitter_document.msg_text.count("):") + twitter_document.msg_text.count(":(") + twitter_document.msg_text.count(":-(") + twitter_document.msg_text.count(":c") + twitter_document.msg_text.count(":[") + 2*twitter_document.msg_text.count("D8") + twitter_document.msg_text.count("D;") + 2*twitter_document.msg_text.count("D=") + twitter_document.msg_text.count("DX");
```

The following emoticons representing smilies were seached for:
`(: , :) , :-) , o) , :] , :3 , :c , :D, C:`

The following emoticons representing sad faces were searched for:
`): , :( , :-( , :c , :[ , D8 , D; , D=, DX`

In addition, the amount of question marks and exclamations were added to each document as features.

This resulted in the following results from the three classifiers:

Decision Tree:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3578               49.4951 %
Incorrectly Classified Instances      3651               50.5049 %
Kappa statistic                          0.254
Mean absolute error                      0.2737
Root mean squared error                  0.4489
Relative absolute error                 79.036  %
Root relative squared error            107.8775 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.716     0.346      0.633     0.716     0.672      0.721    positiv
e
                 0.351     0.13       0.371     0.351     0.361      0.614    negativ
e
                 0.262     0.154      0.316     0.262     0.286      0.572    neutral
                 0.333     0.105      0.364     0.333     0.348      0.633    objecti
ve
Weighted Avg.    0.495     0.229      0.477     0.495     0.484      0.657


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2351  354  358  221 |    a = positive
  461  454  242  136 |    b = negative
  566  291  405  285 |    c = neutral
  336  125  276  368 |    d = objective
```

Naive Bayes:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3459               47.8489 %
Incorrectly Classified Instances      3770               52.1511 %
Kappa statistic                          0.271
Mean absolute error                      0.2747
Root mean squared error                  0.443
Relative absolute error                 79.3219 %
Root relative squared error            106.4743 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.581     0.224      0.684     0.581     0.628      0.73     positiv
e
                 0.5       0.198      0.355     0.5       0.415      0.709    negativ
e
                 0.218     0.124      0.325     0.218     0.261      0.605    neutral
                 0.512     0.165      0.359     0.512     0.422      0.753    objecti
ve
Weighted Avg.    0.478     0.189      0.499     0.478     0.48       0.703


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1909  613  334  428 |    a = positive
  303  646  174  170 |    b = negative
  393  404  338  412 |    c = neutral
  186  159  194  566 |    d = objective
```

SVM:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3773               52.1926 %
Incorrectly Classified Instances      3456               47.8074 %
Kappa statistic                          0.2935
Mean absolute error                      0.3183
Root mean squared error                  0.4041
Relative absolute error                 91.9333 %
Root relative squared error             97.1217 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.732     0.33       0.649     0.732     0.688      0.736    positiv
e
                 0.394     0.107      0.446     0.394     0.419      0.724    negativ
e
                 0.305     0.165      0.335     0.305     0.32       0.586    neutral
                 0.351     0.096      0.398     0.351     0.373      0.724    objecti
ve
Weighted Avg.    0.522     0.219      0.507     0.522     0.513      0.7


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2403  282  391  208 |    a = positive
  426  510  250  107 |    b = negative
  555  249  472  271 |    c = neutral
  321  102  294  388 |    d = objective
```

As you can see this increased the average precision for all classifiers. Most notably, the SVM classifier increased from **49.5% to 50.7%.** This classifier continued to be be the most accurate, correctly classifying **3773** twitter messages or 52.2%.

In trying to continue the improvement of the classifiers, we used senti wordnet to add positive, negative, and objective scores for each document. Iterating through each document, each word was analyzed using senti wordnet and the positive, negative, and objective score for the word (in all of the synsets in which it belongs) was added to to total positive, negative and objective score for the document. This was achieved using the following code:

```
for word in twitter_document.msg_text.split():
    for synset in swn.senti_synsets(word):
        additional_features["posscore"] += synset.pos_score()
        additional_features["negscore"] += synset.neg_score()
        additional_features["objscore"] += synset.obj_score()
```

3 features were added to the arff file: `posscore, negscore, objscore`

The three classifiers then provided the following results with these new features:

Decision Tree:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3613               49.9793 %
Incorrectly Classified Instances      3616               50.0207 %
Kappa statistic                          0.2636
Mean absolute error                      0.2698
Root mean squared error                  0.4552
Relative absolute error                 77.9164 %
Root relative squared error            109.3981 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.715     0.332      0.642     0.715     0.676      0.714    positiv
e
                 0.364     0.125      0.388     0.364     0.376      0.612    negativ
e
                 0.266     0.156      0.317     0.266     0.289      0.565    neutral
                 0.348     0.111      0.362     0.348     0.355      0.627    objecti
ve
Weighted Avg.    0.5       0.224      0.484     0.5       0.49       0.651


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2347  344  356  237 |    a = positive
  442  471  242  138 |    b = negative
  557  277  411  302 |    c = neutral
  312  123  286  384 |    d = objective
```

Naive Bayes:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3419               47.2956 %
Incorrectly Classified Instances      3810               52.7044 %
Kappa statistic                          0.2739
Mean absolute error                      0.2728
Root mean squared error                  0.4476
Relative absolute error                 78.7973 %
Root relative squared error            107.5843 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.548     0.183      0.714     0.548     0.62       0.736    positiv
e
                 0.491     0.191      0.36      0.491     0.415      0.722    negativ
e
                 0.239     0.136      0.323     0.239     0.275      0.601    neutral
                 0.555     0.193      0.342     0.555     0.423      0.757    objecti
ve
Weighted Avg.    0.473     0.176      0.51      0.473     0.479      0.708


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1801  606  362  515 |    a = positive
  242  635  210  206 |    b = negative
  326  390  370  461 |    c = neutral
  155  135  202  613 |    d = objective
```

SVM:

```
=== Stratified cross-validation ===

Correctly Classified Instances        3792               52.4554 %
Incorrectly Classified Instances      3437               47.5446 %
Kappa statistic                          0.2979
Mean absolute error                      0.3175
Root mean squared error                  0.4031
Relative absolute error                 91.7069 %
Root relative squared error             96.8872 %
Total Number of Instances             7229


=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.328      0.65      0.731     0.688      0.738    positiv
e
                 0.398     0.103      0.456     0.398     0.425      0.73     negativ
e
                 0.31      0.164      0.34      0.31      0.324      0.588    neutral
                 0.361     0.098      0.4       0.361     0.38       0.728    objecti
ve
Weighted Avg.    0.525     0.218      0.511     0.525     0.516      0.703


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2400  274  392  218 |    a = positive
  426  514  249  104 |    b = negative
  550  242  479  276 |    c = neutral
  318   98  290  399 |    d = objective
```

Again, we saw an increase in precision and correctly classified instances for all classifiers. Most notably, the SVM classifier increased from **50.7% to 51.1%.** This classifier continued to be be the most accurate, correctly classifying **3792** twitter messages or 52.45%.

With these results we noticed that combining bag of words with counting exclamations, question marks, smile emoticons, sad emoticons, and analyzing the sentiment of each individual word in a Twitter document can in fact increase precision for classifiers. The remaining investigation tested different features and approaches that did not increase precission past 51.1%.