Assignment 2

Natural Language Processing

Ziyi Hu-100981750

Dongyang Li-7996582

Division of Work

The whole assignment was divided equally and we worked together to complete the assignment.

Ziyi Hu:

Part-1

Dongyang Li:

Part-2

Part1:

Train a classifier using the bag-of-words (BOW) representation. This means to use words as features in the arff file. You can eliminate stop words, rare words, punctuation, etc in order to reduce the dimension of the vector space. In fact the file is already in .arff format, with a simple header that you should modify as needed. You can use the Weka filter StringToVector for a first quick experiment.

Step1: First download the arff 2000 sentence file. Save it as the

Saimadata.arff and Saimadata.txt.

Step2: We used the Saimadata.txt file to eliminate stop words,rare

words(only appear one time) and punctuation(use java

program)

Step3: We used Weka to do the classification on the test set. We

used the StringToWord vector filter and three

classifiers(SVM,NB,DT). The classifier is trained on 9

training parts, using the provided class labels to learn

associations between the classes and the data

Before we make any change about the arff file (haven’t eliminate the stop words, rare words and punctuation)

1. For the Naïve Bayes classifier the accuracy for the data is 47.4349%

The result is as following:

=== Summary ===

Correctly Classified Instances 1276 47.4349 %

Incorrectly Classified Instances 1414 52.5651 %

Kappa statistic 0.2542

Mean absolute error 0.1704

Root mean squared error 0.3326

Relative absolute error 88.5092 %

Root relative squared error 107.2504 %

Total Number of Instances 2690

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.373 0.07 0.571 0.373 0.451 0.727 hp

0.277 0.062 0.234 0.277 0.254 0.665 sd

0.417 0.054 0.257 0.417 0.318 0.768 fr

0.261 0.044 0.21 0.261 0.233 0.757 sp

0.597 0.338 0.657 0.597 0.626 0.663 ne

0.209 0.041 0.257 0.209 0.231 0.635 dg

0.436 0.125 0.198 0.436 0.273 0.741 ag

Weighted Avg. 0.474 0.209 0.521 0.474 0.488 0.688

=== Confusion Matrix ===

a b c d e f g <-- classified as

200 31 27 23 170 32 53 | a = hp

18 48 16 7 69 4 11 | b = sd

7 7 48 5 31 3 14 | c = fr

5 8 7 30 48 4 13 | d = sp

100 83 70 68 836 52 191 | e = ne

17 13 9 3 61 36 33 | f = dg

3 15 10 7 57 9 78 | g = ag

2. For the Decision Tree Classifier the accuracy is 50.4089%

The result is as following:

=== Summary ===

Correctly Classified Instances 1356 50.4089 %

Incorrectly Classified Instances 1334 49.5911 %

Kappa statistic 0.1404

Mean absolute error 0.1683

Root mean squared error 0.3341

Relative absolute error 87.4607 %

Root relative squared error 107.7383 %

Total Number of Instances 2690

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.336 0.088 0.486 0.336 0.397 0.676 hp

0.064 0.029 0.133 0.064 0.086 0.541 sd

0.07 0.023 0.121 0.07 0.088 0.521 fr

0.026 0.013 0.081 0.026 0.039 0.527 sp

0.805 0.685 0.56 0.805 0.661 0.56 ne

0.047 0.015 0.17 0.047 0.073 0.55 dg

0.106 0.023 0.25 0.106 0.149 0.59 ag

Weighted Avg. 0.504 0.38 0.433 0.504 0.449 0.58

=== Confusion Matrix ===

a b c d e f g <-- classified as

180 12 5 2 322 6 9 | a = hp

19 11 3 2 128 4 6 | b = sd

16 7 8 2 76 2 4 | c = fr

7 2 4 3 96 1 2 | d = sp

119 40 38 24 1127 20 32 | e = ne

12 7 4 1 136 8 4 | f = dg

17 4 4 3 126 6 19 | g = ag

3. For the SVM the accuracy of the classifier is the 66.5799%

The result is as following:

=== Summary ===

Correctly Classified Instances 1791 66.5799 %

Incorrectly Classified Instances 899 33.4201 %

Kappa statistic 0.4434

Mean absolute error 0.215

Root mean squared error 0.3182

Relative absolute error 111.6807 %

Root relative squared error 102.6121 %

Total Number of Instances 2690

=== Confusion Matrix ===

a b c d e f g <-- classified as

318 8 1 11 192 1 5 | a = hp

22 55 1 4 89 1 1 | b = sd

3 6 38 1 60 2 5 | c = fr

19 0 0 28 67 0 1 | d = sp

74 21 8 17 1254 9 17 | e = ne

10 6 3 1 83 51 18 | f = dg

9 5 3 4 103 8 47 | g = ag

After we eliminate the stop words, rare words and punctuation and that reduce the dimension of the vector space, we expect the result would be better.

We use the getrareword.java to get the list of the words that only appeared once in the text and save the words in the rare word.txt.

Then we use the elimstop.java and elimrare.java to eliminate the stop words and rare words.

Those words can be seen in the rare word.txt and stop word.txt.

We got a new Saimadata elim.arff after doing those things.

We used the weka to training the new data and get the results

1. For the Naïve Bayes classifier the accuracy for the data is 60.855%

The result is as following:

=== Summary ===

Correctly Classified Instances 1637 60.855 %

Incorrectly Classified Instances 1053 39.145 %

Kappa statistic 0.32

Mean absolute error 0.1641

Root mean squared error 0.2838

Relative absolute error 85.2243 %

Root relative squared error 91.5155 %

Total Number of Instances 2690

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.294 0.013 0.844 0.294 0.436 0.789 hp

0.251 0.023 0.426 0.251 0.316 0.764 sd

0.365 0.005 0.764 0.365 0.494 0.773 fr

0.239 0.008 0.571 0.239 0.337 0.783 sp

0.911 0.619 0.615 0.911 0.734 0.737 ne

0.295 0.031 0.398 0.295 0.339 0.795 dg

0.228 0.023 0.418 0.228 0.295 0.788 ag

Weighted Avg. 0.609 0.33 0.626 0.609 0.566 0.76

=== Confusion Matrix ===

a b c d e f g <-- classified as

157 13 3 10 329 13 9 | a = hp

5 43 0 0 117 4 2 | b = sd

1 5 42 1 57 3 6 | c = fr

0 1 0 28 76 8 4 | d = sp

18 26 7 7 1275 36 31 | e = ne

4 6 2 2 103 51 5 | f = dg

1 7 1 1 116 13 41 | g = ag

1. For the Decision Tree Classifier the accuracy is 56.171%

The result is as following:

=== Summary ===

Correctly Classified Instances 1511 56.171 %

Incorrectly Classified Instances 1179 43.829 %

Kappa statistic 0.1334

Mean absolute error 0.1779

Root mean squared error 0.3015

Relative absolute error 92.4192 %

Root relative squared error 97.1981 %

Total Number of Instances 2690

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.262 0.026 0.718 0.262 0.384 0.631 hp

0.023 0.003 0.333 0.023 0.044 0.561 sd

0 0 0 0 0 0.498 fr

0 0 0 0 0 0.55 sp

0.974 0.864 0.55 0.974 0.703 0.562 ne

0.017 0 1 0.017 0.034 0.491 dg

0 0 0 0 0 0.528 ag

Weighted Avg. 0.562 0.455 0.515 0.562 0.447 0.566

=== Confusion Matrix ===

a b c d e f g <-- classified as

140 1 0 0 393 0 0 | a = hp

10 4 0 0 157 0 0 | b = sd

3 0 0 0 112 0 0 | c = fr

0 0 0 0 117 0 0 | d = sp

30 6 0 0 1364 0 0 | e = ne

7 1 0 0 161 3 1 | f = dg

5 0 1 0 174 0 0 | g = ag

1. For the SVM the accuracy of the classifier is the 71.5242%

The result is as following:

=== Summary ===

Correctly Classified Instances 1924 71.5242 %

Incorrectly Classified Instances 766 28.4758 %

Kappa statistic 0.5309

Mean absolute error 0.2133

Root mean squared error 0.316

Relative absolute error 110.8023 %

Root relative squared error 101.8792 %

Total Number of Instances 2690

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.64 0.057 0.737 0.64 0.685 0.818 hp

0.333 0.012 0.655 0.333 0.442 0.755 sd

0.435 0.003 0.847 0.435 0.575 0.793 fr

0.316 0.012 0.544 0.316 0.4 0.798 sp

0.917 0.405 0.711 0.917 0.801 0.762 ne

0.462 0.007 0.816 0.462 0.59 0.782 dg

0.411 0.013 0.692 0.411 0.516 0.751 ag

Weighted Avg. 0.715 0.225 0.716 0.715 0.695 0.776

=== Confusion Matrix ===

a b c d e f g <-- classified as

342 8 0 11 170 1 2 | a = hp

18 57 0 0 93 1 2 | b = sd

7 1 50 2 51 2 2 | c = fr

17 1 0 37 61 0 1 | d = sp

68 10 7 15 1284 4 12 | e = ne

9 2 0 3 65 80 14 | f = dg

3 8 2 0 83 10 74 | g = ag

Discussion:

We trained the classifiers without using list of stop words and rare words. The accuracy is much less than the results we eliminate those words. When we used the Baye the results improved from 47.4349% to 60.855%, when we used the Decision Tree the results increased from 50.4089% to 56.171%, when we used the SVM the results is from 66.5799% to 71.5242%.

So, removing stop words and rear words helped us improve the accuracy in a high level. That’s mainly because those words doesn’t help the machine learn about the emotions in the sentences. There are so many useless words that would obstruct the classifier.

The best training classifier is the SVM. It performed much better than the others in training this data set.

Part2:

Add more features and train more classifiers, in order to try to improve the classification results. For example using the emoticons from the texts as features as should help. Using punctuation marks such as !. !!, !!!, ??, ???, and others elongations could help. Other features can be the number of emotion words in the sentences (you can use lists of emotion words from different resources). Try at least one such resource, for example [WordNet Affect](http://wndomains.fbk.eu/wnaffect.html) (here are the word lists for [download](http://www.site.uottawa.ca/~diana/csi5386/WordNetAffectEmotionLists.zip)). If you use more resources, you can use separate features for number of emotion words for each class that are found in each resource individually. Lists of positive and negative words might also help.

For this part we used emotion words to train the classifiers.

We used the emotion,java to get the data for the .arff file. We counted the number of the emotion words and list them in the sentence. Those emotion words are downloaded from WordNetAffect.

The words would result in how many emotion words are present in each sentence.

We analyzed the file which had already been eliminated useless words by counting emotion words. For every sentence we got a number-plus-sentiment string like “’ 000000 ’ hp”, it represents the numbers of emotional words (anger-disgust-fear-joy-sadness-surprise) in each sentence. Then add this to every sentence at the last as a feature using emotions.java to get part2experiment.arff. Put it to Weka.

We got the result as the following:

1. The accuracy of the Naïve Bayes classifier is : 61.107 %

The result is as following:

=== Summary ===

Correctly Classified Instances 1645 61.107 %

Incorrectly Classified Instances 1047 38.893 %

Kappa statistic 0.359

Mean absolute error 0.1436

Root mean squared error 0.2869

Relative absolute error 74.599 %

Root relative squared error 92.511 %

Total Number of Instances 2692

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.443 0.053 0.673 0.443 0.534 0.79 hp

0.287 0.02 0.495 0.287 0.363 0.754 sd

0.339 0.005 0.736 0.339 0.464 0.745 fr

0.274 0.015 0.451 0.274 0.34 0.777 sp

0.848 0.507 0.645 0.848 0.732 0.75 ne

0.283 0.044 0.308 0.283 0.295 0.787 dg

0.283 0.025 0.443 0.283 0.346 0.78 ag

Weighted Avg. 0.611 0.281 0.601 0.611 0.587 0.764

=== Confusion Matrix ===

a b c d e f g <-- classified as

237 9 2 21 239 20 7 | a = hp

16 49 1 2 98 4 1 | b = sd

8 1 39 1 54 3 9 | c = fr

5 0 0 32 69 6 5 | d = sp

78 22 9 9 1188 61 34 | e = ne

3 9 1 3 100 49 8 | f = dg

5 9 1 3 95 16 51 | g = ag

1. For the Decision Tree Classifier the accuracy is 65.6018 %

The result is as following:

=== Summary ===

Correctly Classified Instances 1766 65.6018 %

Incorrectly Classified Instances 926 34.3982 %

Kappa statistic 0.3846

Mean absolute error 0.1446

Root mean squared error 0.2764

Relative absolute error 75.0954 %

Root relative squared error 89.1063 %

Total Number of Instances 2692

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.437 0.022 0.83 0.437 0.573 0.741 hp

0.211 0.006 0.692 0.211 0.323 0.647 sd

0.278 0.004 0.762 0.278 0.408 0.714 fr

0.205 0.007 0.571 0.205 0.302 0.654 sp

0.965 0.613 0.631 0.965 0.763 0.673 ne

0.254 0.005 0.786 0.254 0.384 0.653 dg

0.244 0.012 0.595 0.244 0.346 0.657 ag

Weighted Avg. 0.656 0.326 0.685 0.656 0.61 0.683

=== Confusion Matrix ===

a b c d e f g <-- classified as

234 3 1 9 283 1 4 | a = hp

11 36 0 1 121 1 1 | b = sd

5 0 32 0 71 4 3 | c = fr

6 0 0 24 87 0 0 | d = sp

18 4 5 7 1352 2 13 | e = ne

5 6 2 0 107 44 9 | f = dg

3 3 2 1 123 4 44 | g = ag

1. For the SVM the accuracy of the classifier is the 72.3254 %

The result is as following:

=== Summary ===

Correctly Classified Instances 1947 72.3254 %

Incorrectly Classified Instances 745 27.6746 %

Kappa statistic 0.5508

Mean absolute error 0.2129

Root mean squared error 0.3153

Relative absolute error 110.5941 %

Root relative squared error 101.6601 %

Total Number of Instances 2692

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.669 0.057 0.746 0.669 0.705 0.83 hp

0.374 0.011 0.696 0.374 0.487 0.762 sd

0.452 0.005 0.788 0.452 0.575 0.791 fr

0.316 0.014 0.507 0.316 0.389 0.8 sp

0.911 0.368 0.729 0.911 0.81 0.779 ne

0.457 0.01 0.76 0.457 0.57 0.823 dg

0.45 0.018 0.643 0.45 0.529 0.776 ag

Weighted Avg. 0.723 0.206 0.719 0.723 0.706 0.792

=== Confusion Matrix ===

a b c d e f g <-- classified as

358 7 1 12 153 1 3 | a = hp

15 64 1 1 87 2 1 | b = sd

6 1 52 2 45 3 6 | c = fr

19 1 1 37 57 1 1 | d = sp

66 11 7 19 1276 6 16 | e = ne

8 3 1 2 62 79 18 | f = dg

8 5 3 0 71 12 81 | g = ag

Discussion:

Compared with previous results, by using Bayes classifier we improved from 60.855% to 61.107%, by Decision Tree improved from 56.171% to 65.6018%, by SVM improved from 71.5242% to 72.3254%. Still the SVM training classifier is the best.

Results.txt in the file shows the best results trained by SVM method.