HW-2 Report

1. **Feature construction:**
   Selected features:
   - Punctuation
   - Bigrams
   - Remove rare features (threshold 5)

   Reason for choosing bigram is increased long-tail specificity of the word so the classifier can easily find out which class has a higher probability, which gives better classifications. Punctuations allows us to identify the polarity of the sentence.

2. **Description of the classifier:**
   For this I have used Naïve Bayes as my classifier. Reason behind choosing this classifier is simple classification based on bayes theorem, it is highly scalable and it is giving more accuracy than other classifier that I used (ex. SVM)

3. **Evaluation technique:**
   This classifier have used 10 fold cross validation.
   Correctly Classified Instances = 495 (47.3231 %)
   Incorrectly Classified Instances = 551 (52.6769 %)
   Kappa statistic = 0.2065
   From confusion matrix we calculated precision, recall and accuracy to evaluate classifier.

4. **Implementation:**
   I have used Tag Helper tool (uses WEKA library) which allow me to choose different features and classifier for sentiment analysis.
   - First removed all empty data, data with garbage sentiment values and in test data replaces empty sentiment with irrelevant data.
   - With its help first I extracted features (punctuation, bigrams, and remove rare features)
   - Then I chose naïve bayes classifier for training. On twitter dataset I trained the classifier.
   - I used 10 fold cross validation. In this partitioned performed randomly. In every 10 cross validation 9 cross used as training and 1 as validation.
   - I have calculated precision, recall and accuracy manually from confusion matrix.
     On Training data set:

     ```
      a   b   c   d   <-- classified as
      19  55  49   6 |   a = irrelevant
      12 186  81  11 |   b = negative
      22 185 259  18 |   c = neutral
       8  60  44  31 |   d = positive
     ```

From this,
Recall for a = 31.148%, b = 38.272%, c = 59.815%, d = 46.97%
Precision for a = 14.729%, b = 64.138%, c = 53.512%, d = 21.678%
Accuracy over all = 47.32%
Kappa = 0.206

On Test data set:

```
   a    b    c    d   <-- classified as
 4196  121    1   12 |    a = irrelevant
  379   35    0    0 |    b = negative
  514   25    5    0 |    c = neutral
  532   53    0    0 |    d = positive
```

From this,
Recall for a = 74.649%, b = 14.957%, c = 83.333%, d = 0%
Precision for a = 96.905%, b = 8.454%, c = 8.454%, d = 0%
Accuracy over all = 72.12%
Kappa = 0.043

5. Analysis of results:

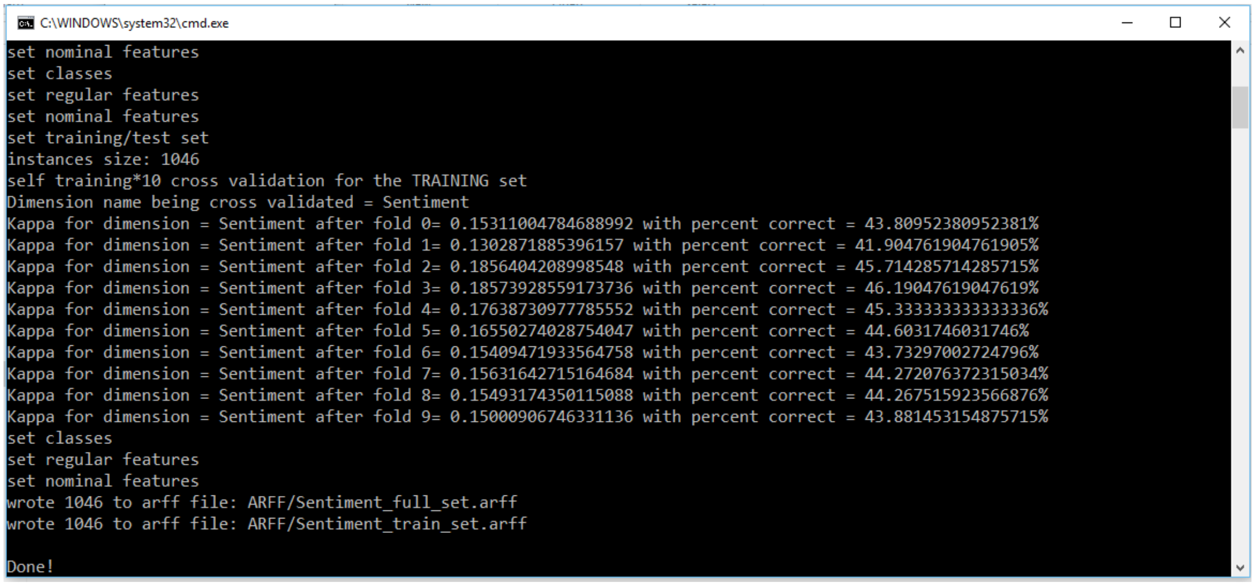
- On Training data set:

```
a   b   c   d   <-- classified as
19  55  49   6 |   a = irrelevant
12 186  81  11 |   b = negative
22 185 259  18 |   c = neutral
8  60  44  31 |   d = positive
```

From this,

Class	Irrelevant	Negative	Neutral	Positive
Precision	14.729%	64.138%	53.512%	21.678%
Recall	31.148%	38.272%	53.512%	21.678%

10 fold cross validation:



C:\WINDOWS\system32\cmd.exe

```
set nominal features
set classes
set regular features
set nominal features
set training/test set
instances size: 1046
self training*10 cross validation for the TRAINING set
Dimension name being cross validated = Sentiment
Kappa for dimension = Sentiment after fold 0= 0.15311004784688992 with percent correct = 43.80952380952381%
Kappa for dimension = Sentiment after fold 1= 0.1302871885396157 with percent correct = 41.904761904761905%
Kappa for dimension = Sentiment after fold 2= 0.1856404208998548 with percent correct = 45.714285714285715%
Kappa for dimension = Sentiment after fold 3= 0.18573928559173736 with percent correct = 46.19047619047619%
Kappa for dimension = Sentiment after fold 4= 0.17638730977785552 with percent correct = 45.333333333333336%
Kappa for dimension = Sentiment after fold 5= 0.16550274028754047 with percent correct = 44.6031746031746%
Kappa for dimension = Sentiment after fold 6= 0.15409471933564758 with percent correct = 43.73297002724796%
Kappa for dimension = Sentiment after fold 7= 0.15631642715164684 with percent correct = 44.272076372315034%
Kappa for dimension = Sentiment after fold 8= 0.15493174350115088 with percent correct = 44.267515923566876%
Kappa for dimension = Sentiment after fold 9= 0.15000906746331136 with percent correct = 43.881453154875715%
set classes
set regular features
set nominal features
wrote 1046 to arff file: ARFF/Sentiment_full_set.arff
wrote 1046 to arff file: ARFF/Sentiment_train_set.arff

Done!
```

6. **Applying classifier to conversational data:**
I have applied classifier to test data and the results are as below:

- On Test data set:

```
a    b    c    d   <-- classified as
4196  121   1   12 |   a = irrelevant
379   35    0    0 |   b = negative
514   25    5    0 |   c = neutral
532   53    0    0 |   d = positive
```

From this,

Class	Irrelevant	Negative	Neutral	Positive
Precision	96.905%	8.454%	8.454%	0%
Recall	74.649%	14.957%	83.333%	0%

Accuracy over all = 72.12%
Kappa = 0.043