Djillali Liabes University of Sidi Bel Abbes
Faculty of Exact Sciences
Department of Computer Sciences

Module : Data Mining
1st year of Master's degeree in
Networks,Systems & Information Security(RSSI)
2021/2022

Text Classification with Weka
TP-04

Student:
Hadjazi Mohammed
Hisham
Group: 01 / RSSI

Module Instructor:
Pr.ELBERRICHI Zakaria
TP Instructor:
Dr.FAHSI.Mahmoud

A paper submitted in fulfilment of the requirements for the
Data Mining TP-04

April 18, 2022

Contents

List of Figures

Chapter 1

Dataset

1.1 Emotion Detection from Text

Emotion detection from text is one of the challenging problems in Natural Language Processing. The reason is the unavailability of labeled dataset and the multi-class nature of the problem. Humans have a variety of emotions and it is difficult to collect enough records for each emotion and hence the problem of class imbalance arises. Here we have a labeled data for emotion detection and the objective is to build an efficient model to detect emotion.

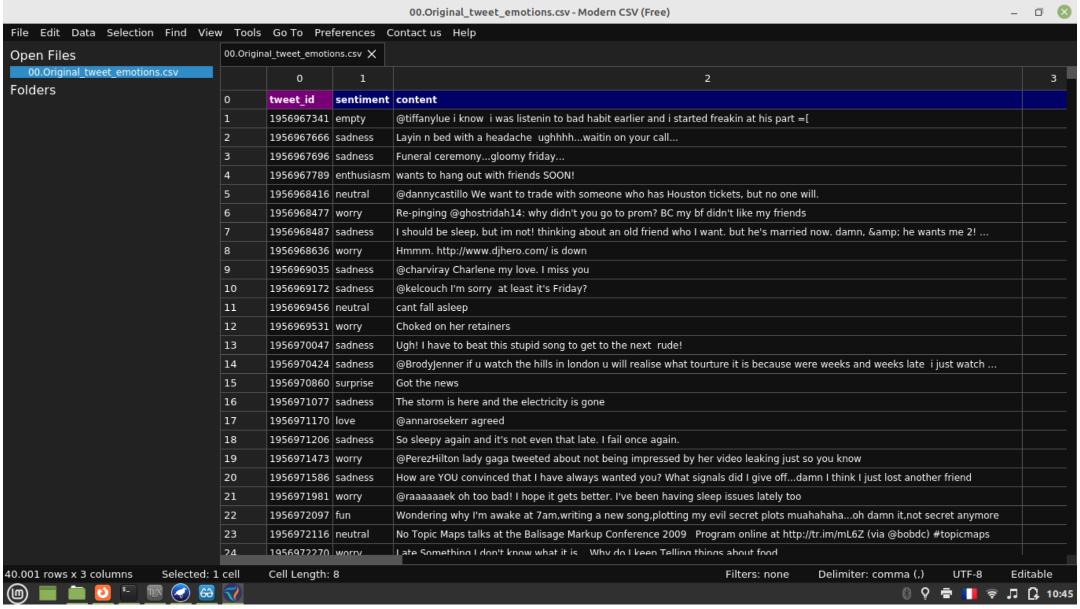
The data is basically a collection of tweets annotated with the emotions behind them. We have three columns tweet id, sentiment, and content. In "content" we have the raw tweet. In "sentiment" we have the emotion behind the tweet. Refer to the starter notebook for more insights.

This public domain dataset is collected from data.world platform. Thanks, data.world for releasing it under Public License.

The data that we have is having 13 different emotion 40000 records. So it's challenging to build an efficient multiclass classification model. We may need to logically reduce the number of classes here and use some advanced methods to build efficient model.

1.1.1 Sample from Dataset

The Original Dataset had 40000 instances and we took only 1000 for our tests which we will see that even 1000 is too much for the specs of my machine.


00.Original_tweet_emotions.csv - Modern CSV (Free)
File Edit Data Selection Find View Tools Go To Preferences Contact us Help
Open Files
00.Original_tweet_emotions.csv
Folders
00.Original_tweet_emotions.csv
10:45

	0	1	2	3
0	tweet_id	sentiment	content	
1	1956967341	empty	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[	
2	1956967666	sadness	Layin n bed with a headache ughhhh...waitin on your call...	
3	1956967696	sadness	Funeral ceremony...gloomy friday...	
4	1956967789	enthusiasm	wants to hang out with friends SOON!	
5	1956968416	neutral	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.	
6	1956968477	worry	Re-pinging @ghostridah14: why didn't you go to prom? BC my bf didn't like my friends	
7	1956968487	sadness	I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, &amp; he wants me 2! ...	
8	1956968636	worry	Hmmm. http://www.djhero.com/ is down	
9	1956969035	sadness	@charviray Charlene my love. I miss you	
10	1956969172	sadness	@kelcouch I'm sorry at least it's Friday?	
11	1956969456	neutral	cant fall asleep	
12	1956969531	worry	Choked on her retainers	
13	1956970047	sadness	Ugh! I have to beat this stupid song to get to the next rude!	
14	1956970424	sadness	@BrodyJenner if u watch the hills in london u will realise what tourture it is because were weeks and weeks late i just watch ...	
15	1956970860	surprise	Got the news	
16	1956971077	sadness	The storm is here and the electricity is gone	
17	1956971170	love	@annarosekerr agreed	
18	1956971206	sadness	So sleepy again and it's not even that late. I fail once again.	
19	1956971473	worry	@PerezHilton lady gaga tweeted about not being impressed by her video leaking just so you know	
20	1956971586	sadness	How are YOU convinced that I have always wanted you? What signals did I give off...damn I think I just lost another friend	
21	1956971981	worry	@raaaaaaek oh too bad! I hope it gets better. I've been having sleep issues lately too	
22	1956972097	fun	Wondering why I'm awake at 7am,writing a new song,plotting my evil secret plots muahahaha...oh damn it,not secret anymore	
23	1956972116	neutral	No Topic Maps talks at the Balisage Markup Conference 2009 Program online at http://tr.im/mL6Z (via @bobdc) #topicmaps	
24	1956972270	worry	Late Something I don't know what it is Why do I keep Telling things about food	

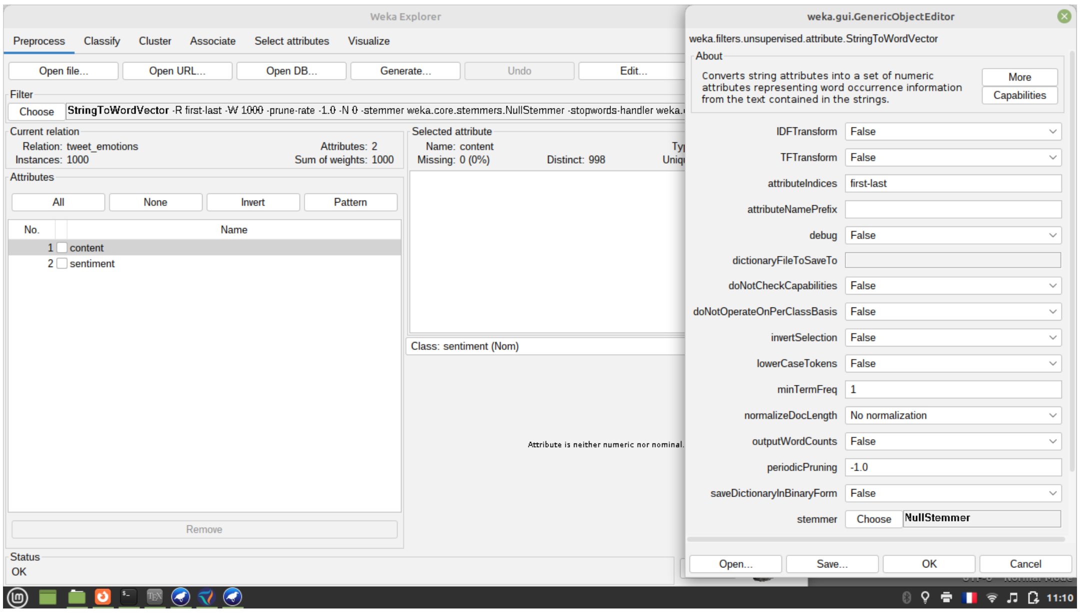
40.001 rows x 3 columns Selected: 1 cell Cell Length: 8 Filters: none Delimiter: comma (,) UTF-8 Editable

FIGURE 1.1: Original CSV

After that we converted the CSV file to an ARFF format.

```
@relation tweet_emotions

@attribute content string
@attribute sentiment {empty,sadness,worry,fun,neutral,hate,enthusiasm,love,surprise,happine,happiness,boredom,relief,anger}

@data
'@tiffanylue i know  i was listenin to bad habit earlier and i started freakin at his part =[',empty
'Layin n bed with a headache  ughhhh...waitin on your call...',sadness
'Funeral ceremony...gloomy friday...',sadness
'wants to hang out with friends SOON!',enthusiasm
'@dannycastillo We want to trade with someone who has Houston tickets  but no one will.',neutral
'Re-pinging @ghostridah14: why didnt you go to prom? BC my bf didnt like my friends',worry
'I should be sleep  but im not! thinking about an old friend who I want. but hes married now. damn  &amp he wants me 2! scandalous!',sadness
'Hmmm. http://www.djhero.com/ is down',worry
'@charviray Charlene my love. I miss you',sadness
'@kelcouch Im sorry  at least its Friday?',sadness
'cant fall asleep',neutral
'Choked on her retainers',worry
'Ugh! I have to beat this stupid song to get to the next  rude!',sadness
'@BrodyJenner if u watch the hills in london u will realise what tourture it is because were weeks and weeks late  i just watch itonlinelol',s
'Got the news',surprise
```

After that we used the StringToWordVector filter to convert our text file to a boolean format that we can use in our classifiers.



Filter: StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.

Current relation
Relation: tweet_emotions Attributes: 2
Instances: 1000 Sum of weights: 1000

Selected attribute
Name: content
Missing: 0 (0%) Distinct: 998

Attributes

No.		Name
1	□	content
2	□	sentiment

Class: sentiment (Nom)

Attribute is neither numeric nor nominal.

Status
OK

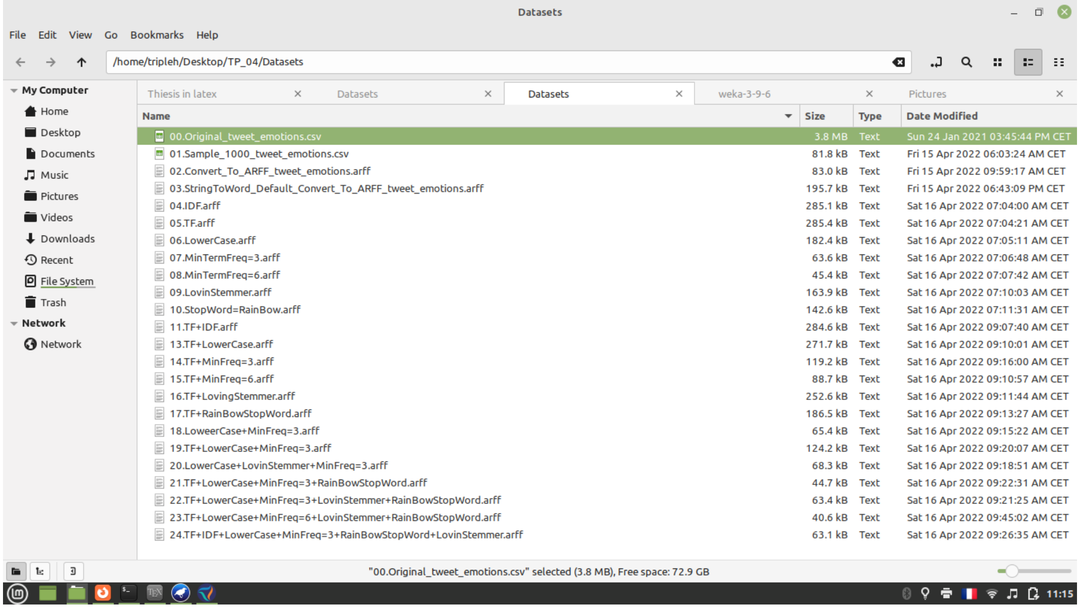
weka.filters.unsupervised.attribute.StringToWordVector

About
Converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings.

IDFTransform	False
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	False
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	Choose NullStemmer

FIGURE 1.2: StringToWordVector

Then we tried to create new datasets with 21 different combinations to test later with the winner classifier.



/home/tripleh/Desktop/TP_04/Datasets

Name	Size	Type	Date Modified
00.Original_tweet_emotions.csv	3.8 MB	Text	Sun 24 Jan 2021 03:45:44 PM CET
01.Sample_1000_tweet_emotions.csv	81.8 kB	Text	Fri 15 Apr 2022 06:03:24 AM CET
02.Convert_To_ARFF_tweet_emotions.arff	83.0 kB	Text	Fri 15 Apr 2022 09:59:17 AM CET
03.StringToWord_Default_Convert_To_ARFF_tweet_emotions.arff	195.7 kB	Text	Fri 15 Apr 2022 06:43:09 PM CET
04.IDF.arff	285.1 kB	Text	Sat 16 Apr 2022 07:04:00 AM CET
05.TF.arff	285.4 kB	Text	Sat 16 Apr 2022 07:04:21 AM CET
06.LowerCase.arff	182.4 kB	Text	Sat 16 Apr 2022 07:05:11 AM CET
07.MinTermFreq=3.arff	63.6 kB	Text	Sat 16 Apr 2022 07:06:48 AM CET
08.MinTermFreq=6.arff	45.4 kB	Text	Sat 16 Apr 2022 07:07:42 AM CET
09.LovinStemmer.arff	163.9 kB	Text	Sat 16 Apr 2022 07:10:03 AM CET
10.StopWord=RainBow.arff	142.6 kB	Text	Sat 16 Apr 2022 07:11:31 AM CET
11.TF+IDF.arff	284.6 kB	Text	Sat 16 Apr 2022 09:07:40 AM CET
13.TF+LowerCase.arff	271.7 kB	Text	Sat 16 Apr 2022 09:10:01 AM CET
14.TF+MinFreq=3.arff	119.2 kB	Text	Sat 16 Apr 2022 09:16:00 AM CET
15.TF+MinFreq=6.arff	88.7 kB	Text	Sat 16 Apr 2022 09:10:57 AM CET
16.TF+LovingStemmer.arff	252.6 kB	Text	Sat 16 Apr 2022 09:11:44 AM CET
17.TF+RainBowStopWord.arff	186.5 kB	Text	Sat 16 Apr 2022 09:13:27 AM CET
18.LoweerCase+MinFreq=3.arff	65.4 kB	Text	Sat 16 Apr 2022 09:15:22 AM CET
19.TF+LowerCase+MinFreq=3.arff	124.2 kB	Text	Sat 16 Apr 2022 09:20:07 AM CET
20.LowerCase+LovinStemmer+MinFreq=3.arff	68.3 kB	Text	Sat 16 Apr 2022 09:18:51 AM CET
21.TF+LowerCase+MinFreq=3+RainBowStopWord.arff	44.7 kB	Text	Sat 16 Apr 2022 09:22:31 AM CET
22.TF+LowerCase+MinFreq=3+LovinStemmer+RainBowStopWord.arff	63.4 kB	Text	Sat 16 Apr 2022 09:21:25 AM CET
23.TF+LowerCase+MinFreq=6+LovinStemmer+RainBowStopWord.arff	40.6 kB	Text	Sat 16 Apr 2022 09:45:02 AM CET
24.TF+IDF+LowerCase+MinFreq=3+RainBowStopWord+LovinStemmer.arff	63.1 kB	Text	Sat 16 Apr 2022 09:26:35 AM CET

"00.Original_tweet_emotions.csv" selected (3.8 MB), Free space: 72.9 GB

FIGURE 1.3: Filter options

Chapter 2

Choosing Algorithms Process

2.1 Introduction

I didn't want to use the default classifiers we were using before, as I wanted to discover and try new classifiers and because Text Classification is more suited to Natural language processing (NLP) than the normal Data Mining Classification methods. So after some research I decided to use the following Classifiers some are our classic classifiers and some are new that I didn't try before.

2.1.1 C4.5

Here we have one of our classic algorithm the Tree base C4.5. The C4.5 algorithm is a famous algorithm of the decision tree which belongs to the data mining filed, but it has been used in many different fields for a long time. However, the C4.5 algorithm is not used in natural language processing (NLP), especially in sentiment classification. We thought that it can be used in the opinion analysis. Therefore, we try applying it into the semantic analysis. This is also very difficult for us to perform it into the sentiment analysis. This is very significantly important for the works and applications in the NLP. From the results which we got, it is true that the C4.5 algorithm is used in the NLP and also in the opinion classification. The aim of this research is to implement the C4.5 algorithm for the emotional analysis of the English documents based on the English sentences of the English training data set. We searched the surveys in the world, which is related to the decision tree, emotional classification. From the below proofs, we found that there is not any research in the world which is similar to this study. We looked for many methodologies to apply the C4.5 algorithm into the sentiment classification for the English documents and then, they are experimented on our data sets. Thus, this proposed model is the originality and novelty research and it also has many meanings in the data mining field, the NLP, the computer science field, etc.[11]

2.1.2 KNN k-nearest neighbors

KNN stands for K Nearest Neighbour. It is a supervised machine learning algorithm that classifies the new text by mapping it with the nearest matches in the training data to make predictions. Since neighbours share similar behavior and characteristics, they can be treated like they belong to the same group. Similarly, the KNN algorithm determines the K nearest neighbours by the closeness and proximity among the training data. The model is trained so that when new data is passed through the model, it can easily match the text to the group or class it belongs to.[7]

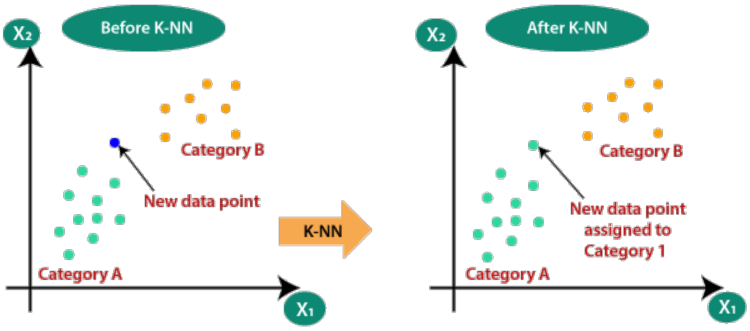


FIGURE 2.1: KNN

In the above image, you can see that new data is assigned to category 1 after passing through the KNN model.[7]

2.1.3 Naive Bayes

Naive Bayes is the simple algorithm that classifies text based on the probability of occurrence of events. This algorithm is based on the Bayes theorem, which helps in finding the conditional probabilities of events that occurred based on the probabilities of occurrence of each individual event.[7]

To understand further how it is used in text classification, let us assume the task is to find whether the given sentence is a statement or a question. Like all machine learning models, this Naive Bayes model also requires a training dataset that contains a collection of sentences labeled with their respective classes. In this case, they are “statement” and “question.” Using the Bayesian equation, the probability is calculated for each class with their respective sentences. Based on the probability value, the algorithm decides whether the sentence belongs to a question class or a statement class.[7]

2.1.4 Multinomial Naive Bayes

Multinomial Naive Bayes is one of the most popular supervised learning classifications that is used for the analysis of the categorical text data.[8]

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.[8]

Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.[8]

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.[13]

2.1.5 Random Forest

The Random Forest (RF) classifiers are suitable for dealing with the high dimensional noisy data in text classification. An RF model comprises a set of decision trees each of which is trained using random subsets of features. Given an instance, the prediction by the RF is obtained via majority voting of the predictions of all the trees in the forest.[3]

Random forests is an averaging ensemble method for classification. The ensemble is a combination of decision trees built from a bootstrap sample from training set. Additionally, in building the decision tree, the split which is chosen when splitting a node is the best split only among a random set of features. This will increase the bias of a single model, but the averaging reduces the variance and can compensate for increase in bias too.

2.1.6 Logistic Regression

This type of statistical analysis (also known as logit model) is often used for predictive analytics and modeling, and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C or D (multinomial regression). It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.[19]

2.1.7 Support Vector Machine (SVN)

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.[16]

Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to have access to a dataset of at most a couple of thousands of tagged samples.[16]

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.[1]

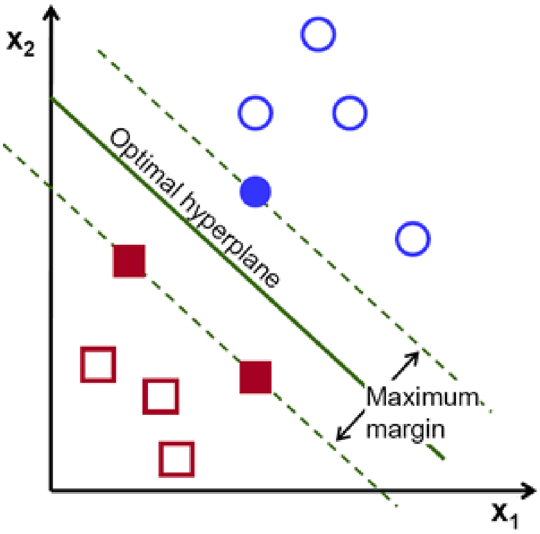


FIGURE 2.2: SVN

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.[1]

Chapter 3

Testing Algorithms

Here we will be using only 2 evaluation tests as the dataset is large with 1000 instances and 3949 attributes, using Leave One Out Fold here will make it impossible for my system to complete the evaluation tasks in reasonable time, A further decrease in the performance as we already made a huge cut from 40,000 instances to a merry 1,000 instance only which is already a big hit to the performance on general. I thought of excluding slow classifiers like C4.5, Random Forest and Logistic Regression, but again these are very commonly used in text classification and replacing them with faster classifiers like One Rule is at least in my perspective a bad decision, So I decided to keep these slow classifiers and sacrifice the number of tests to only 2 in every situation Cross Validation 10 Folds and Percentage Split 66%. leave one out fold combined with Logistic Regression and my huge dataset is a nightmare for my machine at least.

3.1 C4.5 Default Settings

3.1.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         303               30.3    %
Incorrectly Classified Instances       697               69.7    %
Kappa statistic                          0.084
Mean absolute error                      0.115
Root mean squared error                  0.2798
Relative absolute error                 93.977  %
Root relative squared error            113.2393 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.095    0.007    0.222      0.095    0.133      0.134    0.537     0.036     empty
                 0.346    0.214    0.346      0.346    0.346      0.132    0.580     0.301     sadness
                 0.306    0.268    0.315      0.306    0.310      0.038    0.517     0.306     worry
                 0.000    0.004    0.000      0.000    0.000      -0.008   0.530     0.025     fun
                 0.506    0.375    0.293      0.506    0.371      0.113    0.572     0.257     neutral
                 0.130    0.018    0.292      0.130    0.179      0.165    0.565     0.115     hate
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.534     0.016     enthusiasm
                 0.069    0.008    0.200      0.069    0.103      0.102    0.583     0.088     love
                 0.000    0.017    0.000      0.000    0.000      -0.029   0.498     0.049     surprise
                 0.000    0.005    0.000      0.000    0.000      -0.012   0.506     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.508     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.448     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.658     0.006     anger
Weighted Avg.    0.303    0.220    ?          0.303    ?          ?        0.549     0.236

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   2   6   2   0  11   0   0   0   0   0   0   0   0 |   a = empty
   1  85  65   0  78   7   0   6   4   0   0   0   0 |   b = sadness
   2  73  88   1 108   6   1   1   6   2   0   0   0 |   c = worry
   0   6   2   0   6   1   0   0   0   1   0   0   0 |   d = fun
   3  40  63   2 119   2   0   1   4   1   0   0   0 |   e = neutral
   0   7  16   0  22   7   0   0   2   0   0   0   0 |   f = hate
   0   2   6   0   6   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  10   5   0  11   0   0   2   0   1   0   0   0 |   h = love
   1   4  19   0  23   1   0   0   0   0   0   0   0 |   i = surprise
   0   8  10   1  10   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   1   0   4   0   0   0   0   0   0   0   0 |   k = boredom
   0   4   2   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   0   0   3   0   0   0   0   0   0   0   0 |   m = anger
```

3.1.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          89               26.1765 %
Incorrectly Classified Instances       251               73.8235 %
Kappa statistic                          0.0333
Mean absolute error                      0.1204
Root mean squared error                  0.2861
Relative absolute error                 98.0147 %
Root relative squared error            115.0742 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.091    0.006    0.333      0.091    0.143      0.160    0.600     0.060     empty
                 0.277    0.272    0.247      0.277    0.261      0.005    0.515     0.266     sadness
                 0.298    0.276    0.292      0.298    0.295      0.021    0.508     0.274     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.745     0.051     fun
                 0.462    0.366    0.273      0.462    0.343      0.082    0.519     0.250     neutral
                 0.048    0.019    0.143      0.048    0.071      0.049    0.527     0.086     hate
                 0.000    0.000    ?          0.000    ?          ?        0.353     0.012     enthusiasm
                 0.000    0.009    0.000      0.000    0.000      -0.020   0.425     0.038     love
                 0.000    0.012    0.000      0.000    0.000      -0.021   0.452     0.032     surprise
                 0.000    0.003    0.000      0.000    0.000      -0.009   0.435     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.497     0.009     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.376     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.639     0.010     anger
Weighted Avg.    0.262    0.229    ?          0.262    ?          ?        0.508     0.210

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  1  1  3  0  6  0  0  0  0  0  0  0  0 |  a = empty
  0 23 22  0 31  2  0  2  2  1  0  0  0 |  b = sadness
  1 26 28  1 35  3  0  0  0  0  0  0  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  1 18 20  0 36  1  0  1  1  0  0  0  0 |  e = neutral
  0  8  6  0  6  1  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  4  0  3  0  0  0  0  0  0  0  0 |  h = love
  0  1  3  0  8  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  0  0  1  0  0  0  1  0  0  0  0 |  k = boredom
  0  3  1  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

3.2 KNN K=1

3.2.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         275               27.5    %
Incorrectly Classified Instances       725               72.5    %
Kappa statistic                          0.068
Mean absolute error                      0.1183
Root mean squared error                  0.3045
Relative absolute error                 96.6935 %
Root relative squared error            123.2616 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.048    0.018    0.053      0.048    0.050      0.031    0.562     0.029     empty
                 0.309    0.183    0.355      0.309    0.330      0.132    0.576     0.308     sadness
                 0.052    0.056    0.273      0.052    0.087      -0.008   0.529     0.302     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.445     0.014     fun
                 0.779    0.605    0.283      0.779    0.415      0.154    0.601     0.286     neutral
                 0.000    0.008    0.000      0.000    0.000      -0.021   0.487     0.054     hate
                 0.000    0.005    0.000      0.000    0.000      -0.008   0.501     0.014     enthusiasm
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.546     0.046     love
                 0.000    0.013    0.000      0.000    0.000      -0.025   0.499     0.050     surprise
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.532     0.032     happiness
                 0.000    0.011    0.000      0.000    0.000      -0.008   0.328     0.005     boredom
                 0.000    0.011    0.000      0.000    0.000      -0.011   0.393     0.009     relief
                 0.000    0.016    0.000      0.000    0.000      -0.007   0.371     0.003     anger
Weighted Avg.    0.275    0.205    0.234      0.275    0.205      0.064    0.550     0.239

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   1   0   1   0  17   0   0   0   1   0   1   0   0 |   a = empty
   2  76  15   0 138   3   0   1   2   0   3   2   4 |   b = sadness
   9  74  15   0 167   3   3   0   4   0   4   3   6 |   c = worry
   0   1   0   0  14   1   0   0   0   0   0   0   0 |   d = fun
   4  25  10   0 183   0   1   0   3   0   0   5   4 |   e = neutral
   1  13   5   1  29   0   0   0   0   0   2   1   2 |   f = hate
   0   0   3   0  11   0   0   0   0   0   0   0   0 |   g = enthusiasm
   1   9   0   0  18   0   0   0   0   1   0   0   0 |   h = love
   1   9   2   0  35   0   0   0   0   0   1   0   0 |   i = surprise
   0   5   4   0  17   1   1   0   1   0   0   0   0 |   j = happiness
   0   0   0   0   5   0   0   0   1   0   0   0   0 |   k = boredom
   0   2   0   0   9   0   0   0   0   0   0   0   0 |   l = relief
   0   0   0   0   3   0   0   0   0   0   0   0   0 |   m = anger
```

3.2.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          98               28.8235 %
Incorrectly Classified Instances       242               71.1765 %
Kappa statistic                          0.0849
Mean absolute error                      0.1184
Root mean squared error                  0.2977
Relative absolute error                 96.3567 %
Root relative squared error            119.7446 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.021    0.000      0.000    0.000      -0.027   0.667     0.054     empty
                 0.373    0.175    0.408      0.373    0.390      0.205    0.637     0.349     sadness
                 0.043    0.061    0.211      0.043    0.071      -0.036   0.490     0.276     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.281     0.008     fun
                 0.808    0.607    0.284      0.808    0.420      0.177    0.615     0.290     neutral
                 0.000    0.013    0.000      0.000    0.000      -0.028   0.511     0.068     hate
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.699     0.019     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.419     0.037     love
                 0.000    0.024    0.000      0.000    0.000      -0.030   0.509     0.043     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.643     0.048     happiness
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.274     0.007     boredom
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.464     0.014     relief
                 0.000    0.000    ?          0.000    ?          ?        0.418     0.006     anger
Weighted Avg.    0.288    0.201    ?          0.288    ?          ?        0.562     0.239

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  0  0 10  0  0  0  1  0  0  0  0 |  a = empty
  1 31  4  0 44  1  0  0  2  0  0  0  0 |  b = sadness
  3 23  4  0 58  2  0  0  3  0  1  0  0 |  c = worry
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  d = fun
  1  8  2  0 63  0  1  0  2  0  0  1  0 |  e = neutral
  1  4  4  1 11  0  0  0  0  0  0  0  0 |  f = hate
  0  0  3  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  1  6  0  0  7  0  0  0  0  0  0  0  0 |  h = love
  0  1  0  0 11  0  0  0  0  0  0  0  0 |  i = surprise
  0  2  2  0  5  1  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  0  0  4  0  0  0  0  0  0  0  0 |  l = relief
  0  0  0  0  2  0  0  0  0  0  0  0  0 |  m = anger
```

3.3 KNN K=30

3.3.1 Cross Validation (10 Folds)

```
=== Summary ===
Correctly Classified Instances         288               28.8    %
Incorrectly Classified Instances       712               71.2    %
Kappa statistic                          0.0646
Mean absolute error                      0.1229
Root mean squared error                  0.2532
Relative absolute error                100.4386 %
Root relative squared error            102.5088 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.472     0.021     empty
                 0.419    0.318    0.300      0.419    0.350      0.091    0.569     0.314     sadness
                 0.000    0.001    0.000      0.000    0.000      -0.020   0.521     0.321     worry
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.021     fun
                 0.787    0.616    0.282      0.787    0.415      0.153    0.591     0.296     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.549     0.078     hate
                 0.000    0.000    ?          0.000    ?          ?        0.509     0.090     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.516     0.039     love
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.049     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.539     0.037     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.463     0.006     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.649     0.054     relief
                 0.000    0.000    ?          0.000    ?          ?        0.257     0.003     anger
Weighted Avg.    0.288    0.223    ?          0.288    ?          ?        0.549     0.251

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   5   0   0  16   0   0   0   0   0   0   0   0 |   a = empty
   0 103   1   0 142   0   0   0   0   0   0   0   0 |   b = sadness
   0 114   0   0 174   0   0   0   0   0   0   0   0 |   c = worry
   0   3   0   0  13   0   0   0   0   0   0   0   0 |   d = fun
   0  50   0   0 185   0   0   0   0   0   0   0   0 |   e = neutral
   0  20   0   0  34   0   0   0   0   0   0   0   0 |   f = hate
   0   5   0   0   9   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  12   0   0  17   0   0   0   0   0   0   0   0 |   h = love
   0  12   0   0  36   0   0   0   0   0   0   0   0 |   i = surprise
   0  13   0   0  16   0   0   0   0   0   0   0   0 |   j = happiness
   0   2   0   0   4   0   0   0   0   0   0   0   0 |   k = boredom
   0   3   0   0   8   0   0   0   0   0   0   0   0 |   l = relief
   0   1   0   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

3.3.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          84               24.7059 %
Incorrectly Classified Instances       256               75.2941 %
Kappa statistic                          0.0046
Mean absolute error                      0.1227
Root mean squared error                  0.2532
Relative absolute error                 99.8426 %
Root relative squared error            101.8347 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.537     0.036     empty
                 0.988    0.946    0.252      0.988    0.402      0.089    0.537     0.285     sadness
                 0.000    0.004    0.000      0.000    0.000      -0.034   0.485     0.289     worry
                 0.000    0.000    ?          0.000    ?          ?        0.514     0.012     fun
                 0.026    0.046    0.143      0.026    0.043      -0.043   0.454     0.209     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.610     0.117     hate
                 0.000    0.000    ?          0.000    ?          ?        0.784     0.085     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.408     0.035     love
                 0.000    0.000    ?          0.000    ?          ?        0.635     0.051     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.630     0.061     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.613     0.025     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.590     0.217     relief
                 0.000    0.000    ?          0.000    ?          ?        0.530     0.011     anger
Weighted Avg.    0.247    0.242    ?          0.247    ?          ?        0.513     0.215

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0 10  0  0  1  0  0  0  0  0  0  0  0 |  a = empty
  0 82  0  0  1  0  0  0  0  0  0  0  0 |  b = sadness
  0 85  0  0  9  0  0  0  0  0  0  0  0 |  c = worry
  0  2  0  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 75  1  0  2  0  0  0  0  0  0  0  0 |  e = neutral
  0 21  0  0  0  0  0  0  0  0  0  0  0 |  f = hate
  0  4  0  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0 14  0  0  0  0  0  0  0  0  0  0  0 |  h = love
  0 12  0  0  0  0  0  0  0  0  0  0  0 |  i = surprise
  0 10  0  0  0  0  0  0  0  0  0  0  0 |  j = happiness
  0  3  0  0  0  0  0  0  0  0  0  0  0 |  k = boredom
  0  5  0  0  0  0  0  0  0  0  0  0  0 |  l = relief
  0  2  0  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

3.4 Logistic Regression Default Settings

3.4.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         335               33.5    %
Incorrectly Classified Instances       665               66.5    %
Kappa statistic                          0.1154
Mean absolute error                      0.1166
Root mean squared error                  0.2494
Relative absolute error                 95.3049 %
Root relative squared error            100.9371 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.003    0.000      0.000    0.000      -0.008   0.646     0.036     empty
                 0.260    0.111    0.432      0.260    0.325      0.180    0.586     0.340     sadness
                 0.510    0.400    0.340      0.510    0.408      0.101    0.552     0.357     worry
                 0.000    0.004    0.000      0.000    0.000      -0.008   0.440     0.015     fun
                 0.506    0.319    0.328      0.506    0.398      0.165    0.650     0.309     neutral
                 0.056    0.020    0.136      0.056    0.079      0.055    0.580     0.095     hate
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.460     0.018     enthusiasm
                 0.034    0.004    0.200      0.034    0.059      0.072    0.653     0.135     love
                 0.021    0.005    0.167      0.021    0.037      0.043    0.534     0.076     surprise
                 0.000    0.010    0.000      0.000    0.000      -0.017   0.457     0.027     happiness
                 0.000    0.005    0.000      0.000    0.000      -0.006   0.497     0.008     boredom
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.426     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.365     0.003     anger
Weighted Avg.    0.335    0.220    ?          0.335    ?          ?        0.581     0.274

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1   9   0   9   1   0   0   0   0   1   0   0 |   a = empty
   0  64  98   1  72   6   0   2   2   1   0   0   0 |   b = sadness
   0  41 147   2  82   9   0   1   1   4   1   0   0 |   c = worry
   0   2   6   0   6   0   0   0   0   1   0   1   0 |   d = fun
   2  18  88   0 119   3   0   1   1   2   1   0   0 |   e = neutral
   1   8  23   1  17   3   0   0   0   0   1   0   0 |   f = hate
   0   2   6   0   5   0   0   0   0   1   0   0   0 |   g = enthusiasm
   0   4  10   0  12   0   0   1   1   1   0   0   0 |   h = love
   0   2  22   0  23   0   0   0   1   0   0   0   0 |   i = surprise
   0   4  16   0   7   0   1   0   0   0   1   0   0 |   j = happiness
   0   1   2   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   4   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

3.4.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         120               35.2941 %
Incorrectly Classified Instances       220               64.7059 %
Kappa statistic                          0.1651
Mean absolute error                      0.1169
Root mean squared error                  0.2555
Relative absolute error                 95.1918 %
Root relative squared error            102.7788 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.691     0.058     empty
                 0.265    0.082    0.512      0.265    0.349      0.237    0.599     0.372     sadness
                 0.468    0.256    0.411      0.468    0.438      0.204    0.557     0.330     worry
                 0.000    0.006    0.000      0.000    0.000      -0.007   0.668     0.018     fun
                 0.628    0.397    0.320      0.628    0.424      0.195    0.615     0.288     neutral
                 0.000    0.013    0.000      0.000    0.000      -0.028   0.523     0.081     hate
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.342     0.010     enthusiasm
                 0.357    0.021    0.417      0.357    0.385      0.361    0.701     0.246     love
                 0.000    0.021    0.000      0.000    0.000      -0.028   0.565     0.045     surprise
                 0.000    0.024    0.000      0.000    0.000      -0.027   0.589     0.042     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.411     0.010     boredom
                 0.000    0.006    0.000      0.000    0.000      -0.009   0.533     0.027     relief
                 0.000    0.000    ?          0.000    ?          ?        0.343     0.007     anger
Weighted Avg.    0.353    0.185    ?          0.353    ?          ?        0.586     0.269

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  2  0  8  0  0  0  0  1  0  0  0 |  a = empty
  0 22 23  0 29  1  0  5  0  2  0  1  0 |  b = sadness
  1  9 44  1 28  1  0  0  5  5  0  0  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  0  4 19  0 49  2  1  1  1  0  0  1  0 |  e = neutral
  0  7  7  1  6  0  0  0  0  0  0  0  0 |  f = hate
  0  0  2  0  2  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  0  2  0  6  0  0  5  1  0  0  0  0 |  h = love
  0  0  2  0 10  0  0  0  0  0  0  0  0 |  i = surprise
  0  0  3  0  6  0  0  1  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  1  0  3  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

3.5 Naive Bayes Default Settings

3.5.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances          70                7      %
Incorrectly Classified Instances        88                8.8    %
Kappa statistic                          0.1303
Mean absolute error                      0.0974
Root mean squared error                  0.2543
Relative absolute error                537.9313 %
Root relative squared error            277.3066 %
UnClassified Instances                 842               84.2    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 ?        0.013    0.000      ?        ?          ?        0.589     0.025     empty
                 0.000    0.197    0.000      0.000    0.000      -0.039   0.602     0.291     sadness
                 0.558    0.244    0.851      0.558    0.674      0.283    0.341     0.226     worry
                 0.000    0.007    0.000      0.000    0.000      -0.021   0.241     0.010     fun
                 0.182    0.088    0.353      0.182    0.240      0.123    0.513     0.240     neutral
                 1.000    0.076    0.077      1.000    0.143      0.267    0.578     0.063     hate
                 ?        0.006    0.000      ?        ?          ?        0.587     0.017     enthusiasm
                 ?        0.044    0.000      ?        ?          ?        0.585     0.035     love
                 ?        0.051    0.000      ?        ?          ?        0.587     0.057     surprise
                 ?        0.019    0.000      ?        ?          ?        0.587     0.035     happiness
                 ?        0.000    ?          ?        ?          ?        0.591     0.007     boredom
                 ?        0.006    0.000      ?        ?          ?        0.589     0.013     relief
                 ?        0.000    ?          ?        ?          ?        0.591     0.004     anger
Weighted Avg.    0.443    0.195    0.683      0.443    0.533      0.228    0.374     0.215

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = empty
  0  0  0  0  0  0  0  0  1  0  0  0  0 |  b = sadness
  0 22 63  1  8 11  0  5  1  2  0  0  0 |  c = worry
  1  3  0  0  3  0  0  0  2  1  0  0  0 |  d = fun
  1  6 11  0  6  1  1  2  4  0  0  1  0 |  e = neutral
  0  0  0  0  0  1  0  0  0  0  0  0  0 |  f = hate
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  h = love
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  i = surprise
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  k = boredom
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  l = relief
  0  0  0  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

3.5.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         0                0      %
```

```
Kappa statistic                          1
Mean absolute error                    NaN
Root mean squared error                NaN
Relative absolute error                NaN      %
Root relative squared error            NaN      %
UnClassified Instances                 340              100      %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 ?        ?        ?          ?        ?          ?        0.500     0.032     empty
                 ?        ?        ?          ?        ?          ?        0.500     0.244     sadness
                 ?        ?        ?          ?        ?          ?        0.500     0.276     worry
                 ?        ?        ?          ?        ?          ?        0.500     0.009     fun
                 ?        ?        ?          ?        ?          ?        0.500     0.229     neutral
                 ?        ?        ?          ?        ?          ?        0.500     0.062     hate
                 ?        ?        ?          ?        ?          ?        0.500     0.012     enthusiasm
                 ?        ?        ?          ?        ?          ?        0.500     0.041     love
                 ?        ?        ?          ?        ?          ?        0.500     0.035     surprise
                 ?        ?        ?          ?        ?          ?        0.500     0.029     happiness
                 ?        ?        ?          ?        ?          ?        0.500     0.009     boredom
                 ?        ?        ?          ?        ?          ?        0.500     0.015     relief
                 ?        ?        ?          ?        ?          ?        0.500     0.006     anger
Weighted Avg.    ?        ?        ?          ?        ?          ?        ?         ?

=== Confusion Matrix ===

 a b c d e f g h i j k l m   <-- classified as
 0 0 0 0 0 0 0 0 0 0 0 0 0 | a = empty
 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = sadness
 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = worry
 0 0 0 0 0 0 0 0 0 0 0 0 0 | d = fun
 0 0 0 0 0 0 0 0 0 0 0 0 0 | e = neutral
 0 0 0 0 0 0 0 0 0 0 0 0 0 | f = hate
 0 0 0 0 0 0 0 0 0 0 0 0 0 | g = enthusiasm
 0 0 0 0 0 0 0 0 0 0 0 0 0 | h = love
 0 0 0 0 0 0 0 0 0 0 0 0 0 | i = surprise
 0 0 0 0 0 0 0 0 0 0 0 0 0 | j = happiness
 0 0 0 0 0 0 0 0 0 0 0 0 0 | k = boredom
 0 0 0 0 0 0 0 0 0 0 0 0 0 | l = relief
 0 0 0 0 0 0 0 0 0 0 0 0 0 | m = anger
```

3.6 Multinomial Naive Bayes Default Settings

3.6.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         349               34.9    %
Incorrectly Classified Instances       651               65.1    %
Kappa statistic                          0.1157
Mean absolute error                      0.1056
Root mean squared error                  0.265
Relative absolute error                 86.313  %
Root relative squared error            107.255  %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.642     0.039     empty
                 0.402    0.252    0.343      0.402    0.370      0.143    0.606     0.331     sadness
                 0.656    0.492    0.351      0.656    0.457      0.150    0.618     0.391     worry
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.510     0.021     fun
                 0.260    0.125    0.389      0.260    0.311      0.156    0.654     0.343     neutral
                 0.000    0.005    0.000      0.000    0.000      -0.017   0.483     0.059     hate
                 0.000    0.000    ?          0.000    ?          ?        0.475     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.464     0.033     love
                 0.000    0.006    0.000      0.000    0.000      -0.017   0.583     0.060     surprise
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.484     0.029     happiness
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.697     0.016     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.559     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.794     0.020     anger
Weighted Avg.    0.349    0.234    ?          0.349    ?          ?        0.603     0.284

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  11   0   6   0   0   0   0   0   0   0   0 |   a = empty
   0  99 127   0  18   0   0   0   2   0   0   0   0 |   b = sadness
   0  67 189   0  30   1   0   0   0   1   0   0   0 |   c = worry
   0   4   7   0   4   1   0   0   0   0   0   0   0 |   d = fun
   0  59 109   0  61   2   0   0   4   0   0   0   0 |   e = neutral
   0  13  29   2   9   0   0   0   0   0   1   0   0 |   f = hate
   0   2  10   0   2   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  17   8   0   4   0   0   0   0   0   0   0   0 |   h = love
   0  10  26   0  12   0   0   0   0   0   0   0   0 |   i = surprise
   0  11  14   0   4   0   0   0   0   0   0   0   0 |   j = happiness
   0   2   2   0   1   1   0   0   0   0   0   0   0 |   k = boredom
   0   1   6   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

3.6.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         110               32.3529 %
Incorrectly Classified Instances       230               67.6471 %
Kappa statistic                          0.0871
Mean absolute error                      0.107
Root mean squared error                  0.2712
Relative absolute error                 87.0678 %
Root relative squared error            109.0597 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.702     0.118     empty
                 0.337    0.222    0.329      0.337    0.333      0.115    0.592     0.328     sadness
                 0.713    0.557    0.328      0.713    0.450      0.142    0.626     0.414     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.642     0.092     fun
                 0.192    0.122    0.319      0.192    0.240      0.085    0.635     0.325     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.438     0.061     hate
                 0.000    0.000    ?          0.000    ?          ?        0.349     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.483     0.049     love
                 0.000    0.009    0.000      0.000    0.000      -0.018   0.704     0.079     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.645     0.050     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.757     0.030     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.515     0.021     relief
                 0.000    0.000    ?          0.000    ?          ?        0.713     0.019     anger
Weighted Avg.    0.324    0.236    ?          0.324    ?          ?        0.605     0.285

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  7  0  4  0  0  0  0  0  0  0  0 |  a = empty
  0 28 47  0  7  0  0  0  1  0  0  0  0 |  b = sadness
  0 20 67  0  7  0  0  0  0  0  0  0  0 |  c = worry
  0  1  2  0  0  0  0  0  0  0  0  0  0 |  d = fun
  0 14 47  0 15  0  0  0  2  0  0  0  0 |  e = neutral
  0  4 12  1  4  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  4  0  2  0  0  0  0  0  0  0  0 |  h = love
  0  3  6  0  3  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  0  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  3  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

3.7 Random Forest Default Settings

3.7.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         326               32.6    %
Incorrectly Classified Instances       674               67.4    %
Kappa statistic                          0.0968
Mean absolute error                      0.1177
Root mean squared error                  0.2465
Relative absolute error                 96.1726 %
Root relative squared error             99.7702 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.034     empty
                 0.378    0.252    0.329      0.378    0.352      0.121    0.592     0.356     sadness
                 0.372    0.289    0.342      0.372    0.356      0.080    0.586     0.342     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.570     0.032     fun
                 0.536    0.357    0.316      0.536    0.397      0.155    0.636     0.328     neutral
                 0.000    0.003    0.000      0.000    0.000      -0.013   0.587     0.079     hate
                 0.000    0.000    ?          0.000    ?          ?        0.562     0.018     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.055     love
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.581     0.061     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.533     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.337     0.006     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.485     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.527     0.004     anger
Weighted Avg.    0.326    0.229    ?          0.326    ?          ?        0.595     0.275

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   5   5   0  11   0   0   0   0   0   0   0   0 |   a = empty
   0  93  72   0  79   2   0   0   0   0   0   0   0 |   b = sadness
   0  82 107   0  99   0   0   0   0   0   0   0   0 |   c = worry
   0   2   3   0  10   1   0   0   0   0   0   0   0 |   d = fun
   0  46  62   0 126   0   0   0   1   0   0   0   0 |   e = neutral
   0  15  21   1  17   0   0   0   0   0   0   0   0 |   f = hate
   0   2   7   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0   9  11   0   9   0   0   0   0   0   0   0   0 |   h = love
   0  10   9   0  29   0   0   0   0   0   0   0   0 |   i = surprise
   0  15  11   0   3   0   0   0   0   0   0   0   0 |   j = happiness
   0   2   1   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   3   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

3.7.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         110               32.3529 %
Incorrectly Classified Instances       230               67.6471 %
Kappa statistic                          0.1022
Mean absolute error                      0.118
Root mean squared error                  0.247
Relative absolute error                 96.0238 %
Root relative squared error             99.3641 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.625     0.056     empty
                 0.434    0.296    0.321      0.434    0.369      0.126    0.604     0.341     sadness
                 0.372    0.252    0.361      0.372    0.366      0.119    0.627     0.363     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.710     0.064     fun
                 0.500    0.344    0.302      0.500    0.377      0.136    0.616     0.332     neutral
                 0.000    0.003    0.000      0.000    0.000      -0.014   0.595     0.088     hate
                 0.000    0.000    ?          0.000    ?          ?        0.731     0.046     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.683     0.103     love
                 0.000    0.000    ?          0.000    ?          ?        0.696     0.098     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.478     0.028     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.485     0.009     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.504     0.018     relief
                 0.000    0.000    ?          0.000    ?          ?        0.843     0.020     anger
Weighted Avg.    0.324    0.221    ?          0.324    ?          ?        0.618     0.277

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  3  1  0  7  0  0  0  0  0  0  0  0 |  a = empty
  0 36 19  0 27  1  0  0  0  0  0  0  0 |  b = sadness
  0 31 35  0 28  0  0  0  0  0  0  0  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  0 20 19  0 39  0  0  0  0  0  0  0  0 |  e = neutral
  0  7 10  1  3  0  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  9  4  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  1  1  0 10  0  0  0  0  0  0  0  0 |  i = surprise
  0  3  3  0  4  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  1  0  3  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

3.8 Support Vector Machine Default Settings

3.8.1 Cross Validation (10 Folds)

```
=== Summary ===
Correctly Classified Instances         337               33.7    %
Incorrectly Classified Instances       663               66.3    %
Kappa statistic                          0.1206
Mean absolute error                      0.1341
Root mean squared error                  0.2556
Relative absolute error                109.572  %
Root relative squared error            103.4627 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.048    0.006    0.143      0.048    0.071      0.071    0.727     0.049     empty
                 0.455    0.272    0.353      0.455    0.398      0.170    0.606     0.311     sadness
                 0.406    0.295    0.358      0.406    0.380      0.107    0.576     0.330     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.560     0.022     fun
                 0.434    0.267    0.333      0.434    0.377      0.154    0.621     0.299     neutral
                 0.019    0.011    0.091      0.019    0.031      0.017    0.574     0.095     hate
                 0.000    0.002    0.000      0.000    0.000      -0.005   0.402     0.013     enthusiasm
                 0.034    0.004    0.200      0.034    0.059      0.072    0.610     0.112     love
                 0.021    0.020    0.050      0.021    0.029      0.001    0.555     0.054     surprise
                 0.069    0.002    0.500      0.069    0.121      0.178    0.522     0.073     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.197     0.005     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.534     0.013     relief
                 0.000    0.000    ?          0.000    ?          ?        0.281     0.003     anger
Weighted Avg.    0.337    0.216    ?          0.337    ?          ?        0.589     0.257

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   1   7   5   0   6   0   0   1   1   0   0   0   0 |   a = empty
   2 112  69   0  57   3   0   0   3   0   0   0   0 |   b = sadness
   2  92 117   0  69   3   1   0   4   0   0   0   0 |   c = worry
   0   4   6   0   5   1   0   0   0   0   0   0   0 |   d = fun
   1  53  69   0 102   2   1   0   7   0   0   0   0 |   e = neutral
   0  15  21   1  15   1   0   0   0   1   0   0   0 |   f = hate
   0   3   6   0   4   0   0   1   0   0   0   0   0 |   g = enthusiasm
   0  11   4   0   9   1   0   1   2   1   0   0   0 |   h = love
   1   7  16   0  22   0   0   1   1   0   0   0   0 |   i = surprise
   0  11   9   0   5   0   0   1   1   2   0   0   0 |   j = happiness
   0   0   2   0   3   0   0   0   1   0   0   0   0 |   k = boredom
   0   2   2   0   7   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

3.8.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         108               31.7647 %
Incorrectly Classified Instances       232               68.2353 %
Kappa statistic                          0.1021
Mean absolute error                      0.1343
Root mean squared error                  0.256
Relative absolute error                109.3011 %
Root relative squared error            102.9845 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.738     0.082     empty
                 0.482    0.304    0.339      0.482    0.398      0.161    0.580     0.286     sadness
                 0.404    0.293    0.345      0.404    0.373      0.107    0.576     0.317     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.540     0.016     fun
                 0.346    0.260    0.284      0.346    0.312      0.081    0.599     0.274     neutral
                 0.000    0.013    0.000      0.000    0.000      -0.028   0.563     0.143     hate
                 0.250    0.003    0.500      0.250    0.333      0.348    0.682     0.138     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.458     0.055     love
                 0.167    0.018    0.250      0.167    0.200      0.181    0.616     0.081     surprise
                 0.000    0.003    0.000      0.000    0.000      -0.009   0.572     0.034     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.282     0.008     boredom
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.610     0.031     relief
                 0.000    0.000    ?          0.000    ?          ?        0.771     0.013     anger
Weighted Avg.    0.318    0.216    ?          0.318    ?          ?        0.583     0.240

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  4  2  0  5  0  0  0  0  0  0  0  0 |  a = empty
  0 40 19  0 21  2  0  0  1  0  0  0  0 |  b = sadness
  0 30 38  0 23  1  0  0  2  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 21 24  0 27  1  1  0  2  1  0  1  0 |  e = neutral
  0  7  8  1  4  0  0  0  1  0  0  0  0 |  f = hate
  0  1  1  0  1  0  1  0  0  0  0  0  0 |  g = enthusiasm
```

```
 0  5  6  0  3  0  0  0  0  0  0  0  0 |  h = love
 0  1  4  0  5  0  0  0  2  0  0  0  0 |  i = surprise
 0  4  4  0  2  0  0  0  0  0  0  0  0 |  j = happiness
 0  2  1  0  0  0  0  0  0  0  0  0  0 |  k = boredom
 0  3  0  0  2  0  0  0  0  0  0  0  0 |  l = relief
 0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

3.9 Conclusion

Correctly Classified Instances by Algorithm				
Evaluation Process	Cross Validation 10 Folds	Percentage Split 66%	AVG Algorithms	Rank Algorithms
KNN K=1	27.5000%	28.8235%	28.1618%	6
KNN k=30	28.8000%	24.7059%	26.7530%	7
Naïve Bayes	7.0000%	0.0000%	3.5000%	8
Multinomial Naïve Bayes	34.9000%	32.3529%	33.6265%	2
C4.5	30.3000%	26.1765%	28.2383%	5
Random Forest	32.6000%	32.3529%	32.4765%	4
Logistic Regression	33.5000%	35.2941%	34.3971%	1
SVN	33.7000%	31.7647%	32.7324%	3

Here we see that most algorithms are very close to each other excluding Naive Bayes which was the worst performer and even getting 0% in the Percentage Split 66%. Best performance was from Logistic Regression with 35.2941% correctly classified instances in the Percentage Split 66% Second place was for Multinomial Naive Bayes with a result of 34.9% in the Cross Validation 10 Folds test. Third place was for Support Vector Machine with 33.7% in the cross validation 10 folds test. However I have decided to use Multinomial Naive Bayes for the rest of the tests because Logistic Regression is very slow on my machine specially with the size of my dataset.

Chapter 4

Testing using Multinomial Naive Bayes

As we said in Chapter 3 we will be using Multinomial Naive Bayes because of its very fast execution time even on large datasets and its great performance as it was second only to Logistic Regression and by a small margin.

4.1 IDF

While computing TF, all terms are considered equally important. However it is known that certain terms, such as “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.[2]

IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as “is” is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.[2]

$$idf(t) = N/df$$

Now there are few other problems with the IDF , in case of a large corpus,say 100,000,000 , the IDF value explodes , to avoid the effect we take the log of idf .[2]

During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.[2]

that’s the final formula:

Formula :

$$idf(t) = log(N/(df+1))$$

tf-idf now is a the right measure to evaluate how important a word is to a document in a collection or corpus.here are many different variations of TF-IDF but for now let us concentrate on the this basic version.

Formula :

$$tf-idf(t,d) = tf(t,d) * log(N/(df+1))$$

4.1.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         168               16.8    %
Incorrectly Classified Instances       832               83.2    %
Kappa statistic                          0.0521
Mean absolute error                      0.1281
Root mean squared error                  0.3401
Relative absolute error                104.7022 %
Root relative squared error            137.6793 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.048    0.047    0.021      0.048    0.029      0.000    0.599     0.031     empty
                 0.236    0.135    0.363      0.236    0.286      0.118    0.615     0.323     sadness
                 0.212    0.115    0.427      0.212    0.283      0.125    0.554     0.353     worry
                 0.063    0.032    0.031      0.063    0.042      0.022    0.432     0.022     fun
                 0.094    0.106    0.214      0.094    0.130      -0.017   0.578     0.278     neutral
                 0.148    0.079    0.096      0.148    0.117      0.056    0.509     0.064     hate
                 0.071    0.077    0.013      0.071    0.022      -0.002   0.518     0.019     enthusiasm
                 0.103    0.077    0.038      0.103    0.056      0.016    0.568     0.042     love
                 0.167    0.107    0.073      0.167    0.101      0.041    0.586     0.064     surprise
                 0.172    0.077    0.063      0.172    0.092      0.059    0.564     0.089     happiness
                 0.000    0.027    0.000      0.000    0.000      -0.013   0.518     0.010     boredom
                 0.000    0.037    0.000      0.000    0.000      -0.021   0.517     0.012     relief
                 0.000    0.023    0.000      0.000    0.000      -0.008   0.758     0.016     anger
Weighted Avg.    0.168    0.108    0.275      0.168    0.199      0.068    0.573     0.258

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  1  3  3  0  4  2  2  1  2  0  1  1  1 |  a = empty
  9 58 34  9 21 15 15 22 27 22  2  6  6 |  b = sadness
 14 44 61  6 31 30 24 19 23 15  7  9  5 |  c = worry
  1  2  2  1  1  3  0  0  3  2  0  1  0 |  d = fun
 16 27 31  7 22 17 18 16 32 18 11 15  5 |  e = neutral
  2  6  7  3  4  8  8  5  5  3  2  0  1 |  f = hate
  0  2  2  0  2  2  1  1  0  4  0  0  0 |  g = enthusiasm
  1  7  0  1  2  1  1  3  5  5  0  2  1 |  h = love
  1  6  1  2  9  4  2  6  8  4  1  1  3 |  i = surprise
  1  4  1  0  3  0  6  4  2  5  1  2  0 |  j = happiness
  1  1  0  0  2  0  0  0  1  0  0  0  1 |  k = boredom
  0  0  1  3  2  0  0  1  1  2  1  0  0 |  l = relief
  0  0  0  0  0  1  0  0  1  0  1  0  0 |  m = anger
```

4.1.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          71               20.8824 %
Incorrectly Classified Instances       269               79.1176 %
Kappa statistic                          0.0894
Mean absolute error                      0.122
Root mean squared error                  0.3279
Relative absolute error                 99.3386 %
Root relative squared error            131.9021 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.182    0.027    0.182      0.182    0.182      0.154    0.691     0.090     empty
                 0.229    0.121    0.380      0.229    0.286      0.131    0.597     0.320     sadness
                 0.287    0.150    0.422      0.287    0.342      0.157    0.572     0.372     worry
                 0.333    0.027    0.100      0.333    0.154      0.170    0.481     0.173     fun
                 0.141    0.107    0.282      0.141    0.188      0.045    0.594     0.283     neutral
                 0.095    0.063    0.091      0.095    0.093      0.032    0.514     0.100     hate
                 0.000    0.083    0.000      0.000    0.000      -0.033   0.461     0.016     enthusiasm
                 0.214    0.083    0.100      0.214    0.136      0.092    0.562     0.073     love
                 0.250    0.101    0.083      0.250    0.125      0.090    0.748     0.089     surprise
                 0.300    0.085    0.097      0.300    0.146      0.126    0.690     0.091     happiness
                 0.000    0.027    0.000      0.000    0.000      -0.016   0.624     0.015     boredom
                 0.000    0.021    0.000      0.000    0.000      -0.018   0.454     0.016     relief
                 0.000    0.009    0.000      0.000    0.000      -0.007   0.624     0.011     anger
Weighted Avg.    0.209    0.112    0.296      0.209    0.235      0.104    0.589     0.266

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  2  1  2  0  2  0  1  0  2  0  0  0  1 |  a = empty
  2 19 13  1  9  3  4  6 10  9  4  2  1 |  b = sadness
  3 14 27  4  9  5  7  4  9  9  2  0  1 |  c = worry
  0  0  0  1  0  0  0  0  2  0  0  0  0 |  d = fun
  1  8 15  2 11  8  9  9  6  5  1  3  0 |  e = neutral
  1  1  4  1  2  2  1  3  3  2  1  0  0 |  f = hate
  0  2  1  0  0  0  0  0  0  1  0  0  0 |  g = enthusiasm
  1  3  1  1  2  0  2  3  0  1  0  0  0 |  h = love
  0  1  1  0  1  2  0  2  3  0  1  1  0 |  i = surprise
  0  0  0  0  1  1  4  1  0  3  0  0  0 |  j = happiness
  1  0  0  0  0  1  0  0  0  0  0  1  0 |  k = boredom
  0  1  0  0  1  0  0  2  0  1  0  0  0 |  l = relief
  0  0  0  0  1  0  0  0  1  0  0  0  0 |  m = anger
```

4.2 TF

Suppose we have a set of English text documents and wish to rank which document is most relevant to the query , “Data Science is awesome !” A simple way to start out is by eliminating documents that do not contain all three words “Data”,”is”, “Science”, and “awesome”, but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency.[2]

The weight of a term that occurs in a document is simply proportional to the term frequency.

Formula :

$$tf(t,d) = countoftind/numberofwordsind$$

4.2.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         342               34.2    %
Incorrectly Classified Instances       658               65.8    %
Kappa statistic                          0.0982
Mean absolute error                      0.1075
Root mean squared error                  0.257 
Relative absolute error                 87.8475 %
Root relative squared error            104.0264 %
Total Number of Instances             1000     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.636     0.038     empty
                 0.350    0.210    0.352      0.350    0.351      0.140    0.605     0.337     sadness
                 0.729    0.596    0.331      0.729    0.456      0.126    0.624     0.391     worry
                 0.000    0.000    ?          0.000    ?          ?        0.516     0.020     fun
                 0.196    0.093    0.393      0.196    0.261      0.136    0.658     0.347     neutral
                 0.000    0.003    0.000      0.000    0.000      -0.013   0.474     0.062     hate
                 0.000    0.000    ?          0.000    ?          ?        0.476     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.451     0.032     love
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.580     0.059     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.467     0.027     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.704     0.017     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.556     0.014     relief
                 0.000    0.000    ?          0.000    ?          ?        0.794     0.018     anger
Weighted Avg.    0.342    0.245    ?          0.342    ?          ?        0.604     0.287     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   5  12   0   4   0   0   0   0   0   0   0   0 |   a = empty
   0  86 147   0  13   0   0   0   0   0   0   0   0 |   b = sadness
   0  57 210   0  20   1   0   0   0   0   0   0   0 |   c = worry
   0   2   9   0   4   1   0   0   0   0   0   0   0 |   d = fun
   0  50 138   0  46   0   0   0   1   0   0   0   0 |   e = neutral
   1  11  34   0   8   0   0   0   0   0   0   0   0 |   f = hate
   0   1  12   0   1   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  12  13   0   4   0   0   0   0   0   0   0   0 |   h = love
   0   9  29   0  10   0   0   0   0   0   0   0   0 |   i = surprise
   0   9  18   0   2   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   3   0   1   1   0   0   0   0   0   0   0 |   k = boredom
   0   1   8   0   2   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

4.2.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         103               30.2941 %
Incorrectly Classified Instances       237               69.7059 %
Kappa statistic                          0.0516
Mean absolute error                      0.1086
Root mean squared error                  0.2612
Relative absolute error                 88.4163 %
Root relative squared error            105.0707 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.699     0.112     empty
                 0.241    0.175    0.308      0.241    0.270      0.072    0.586     0.332     sadness
                 0.777    0.683    0.303      0.777    0.436      0.092    0.636     0.418     worry
                 0.000    0.000    ?          0.000    ?          ?        0.644     0.046     fun
                 0.128    0.088    0.303      0.128    0.180      0.057    0.645     0.331     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.431     0.062     hate
                 0.000    0.000    ?          0.000    ?          ?        0.363     0.016     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.472     0.047     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.698     0.074     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.622     0.044     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.754     0.030     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.519     0.022     relief
                 0.000    0.000    ?          0.000    ?          ?        0.725     0.021     anger
Weighted Avg.    0.303    0.252    ?          0.303    ?          ?        0.607     0.287

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  7  0  4  0  0  0  0  0  0  0  0 |  a = empty
  0 20 57  0  6  0  0  0  0  0  0  0  0 |  b = sadness
  0 15 73  0  6  0  0  0  0  0  0  0  0 |  c = worry
  0  0  3  0  0  0  0  0  0  0  0  0  0 |  d = fun
  0 12 55  0 10  0  0  0  1  0  0  0  0 |  e = neutral
  0  4 16  0  1  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  6  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  2  8  0  2  0  0  0  0  0  0  0  0 |  i = surprise
  0  3  7  0  0  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  3  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.3 Lower Case

This option will make all the words in lower case form. this will help make the the same words considered the same.

4.3.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         348               34.8    %
Incorrectly Classified Instances       652               65.2    %
Kappa statistic                          0.1121
Mean absolute error                      0.1045
Root mean squared error                  0.2643
Relative absolute error                 85.4175 %
Root relative squared error            106.9625 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.649     0.050     empty
                 0.411    0.265    0.336      0.411    0.369      0.136    0.616     0.335     sadness
                 0.663    0.487    0.355      0.663    0.462      0.160    0.632     0.400     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.501     0.022     fun
                 0.238    0.129    0.361      0.238    0.287      0.128    0.665     0.331     neutral
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.493     0.063     hate
                 0.000    0.000    ?          0.000    ?          ?        0.479     0.016     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.467     0.035     love
                 0.000    0.002    0.000      0.000    0.000      -0.010   0.602     0.064     surprise
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.440     0.025     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.573     0.010     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.563     0.017     relief
                 0.000    0.000    ?          0.000    ?          ?        0.778     0.024     anger
Weighted Avg.    0.348    0.236    ?          0.348    ?          ?        0.612     0.286

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  13   0   4   0   0   0   0   0   0   0   0 |   a = empty
   0 101 122   0  23   0   0   0   0   0   0   0   0 |   b = sadness
   0  71 191   0  25   0   0   0   0   1   0   0   0 |   c = worry
   0   5   6   0   4   1   0   0   0   0   0   0   0 |   d = fun
   0  61 117   0  56   0   0   0   1   0   0   0   0 |   e = neutral
   0  16  26   1  10   0   0   0   1   0   0   0   0 |   f = hate
   0   3   8   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  17   6   0   6   0   0   0   0   0   0   0   0 |   h = love
   0  11  22   0  14   1   0   0   0   0   0   0   0 |   i = surprise
   0  10  16   0   3   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   5   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.3.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         111               32.6471 %
Incorrectly Classified Instances       229               67.3529 %
Kappa statistic                          0.0893
Mean absolute error                      0.1067
Root mean squared error                  0.2716
Relative absolute error                 86.8806 %
Root relative squared error            109.2371 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.720     0.157     empty
                 0.325    0.206    0.338      0.325    0.331      0.121    0.614     0.325     sadness
                 0.713    0.573    0.322      0.713    0.444      0.128    0.619     0.415     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.662     0.177     fun
                 0.218    0.126    0.340      0.218    0.266      0.109    0.634     0.329     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.484     0.069     hate
                 0.000    0.000    ?          0.000    ?          ?        0.373     0.018     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.481     0.053     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.730     0.082     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.529     0.033     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.769     0.038     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.529     0.024     relief
                 0.000    0.000    ?          0.000    ?          ?        0.695     0.017     anger
Weighted Avg.    0.326    0.238    ?          0.326    ?          ?        0.610     0.287

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  8  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 27 49  0  7  0  0  0  0  0  0  0  0 |  b = sadness
  0 18 67  0  9  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 14 46  0 17  0  0  0  1  0  0  0  0 |  e = neutral
  0  4 13  1  3  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  6  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  3  4  0  5  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  3  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.4 Minimum Frequency = 3

Words will be considered only if they are repeated 3 times at least.

4.4.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         361               36.1    %
Incorrectly Classified Instances       639               63.9    %
Kappa statistic                          0.1428
Mean absolute error                      0.1086
Root mean squared error                  0.2505
Relative absolute error                 88.729  %
Root relative squared error            101.3927 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.007   0.615     0.032     empty
                 0.435    0.221    0.391      0.435    0.412      0.206    0.652     0.405     sadness
                 0.531    0.360    0.374      0.531    0.439      0.158    0.636     0.421     worry
                 0.000    0.000    ?          0.000    ?          ?        0.520     0.019     fun
                 0.430    0.247    0.348      0.430    0.385      0.171    0.653     0.350     neutral
                 0.000    0.013    0.000      0.000    0.000      -0.026   0.577     0.077     hate
                 0.000    0.000    ?          0.000    ?          ?        0.465     0.014     enthusiasm
                 0.000    0.005    0.000      0.000    0.000      -0.012   0.525     0.042     love
                 0.000    0.006    0.000      0.000    0.000      -0.017   0.554     0.061     surprise
                 0.000    0.002    0.000      0.000    0.000      -0.008   0.557     0.035     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.587     0.009     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.453     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.681     0.009     anger
Weighted Avg.    0.361    0.217    ?          0.361    ?          ?        0.624     0.314

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   2  12   0   7   0   0   0   0   0   0   0   0 |   a = empty
   1 107  82   0  47   5   0   2   2   0   0   0   0 |   b = sadness
   0  65 153   0  62   6   0   1   1   0   0   0   0 |   c = worry
   0   3   6   0   6   0   0   0   0   1   0   0   0 |   d = fun
   1  39  90   0 101   1   0   0   2   1   0   0   0 |   e = neutral
   0  17  23   0  14   0   0   0   0   0   0   0   0 |   f = hate
   0   2   7   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  13   6   0   9   0   0   0   1   0   0   0   0 |   h = love
   0  11  15   0  21   0   0   1   0   0   0   0   0 |   i = surprise
   0  10   9   0   9   0   0   1   0   0   0   0   0 |   j = happiness
   0   3   0   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   4   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.4.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         105               30.8824 %
Incorrectly Classified Instances       235               69.1176 %
Kappa statistic                          0.082
Mean absolute error                      0.1114
Root mean squared error                  0.2575
Relative absolute error                 90.704  %
Root relative squared error            103.5794 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.654     0.051     empty
                 0.373    0.257    0.320      0.373    0.344      0.111    0.620     0.380     sadness
                 0.489    0.394    0.322      0.489    0.388      0.086    0.613     0.423     worry
                 0.000    0.000    ?          0.000    ?          ?        0.641     0.018     fun
                 0.359    0.229    0.318      0.359    0.337      0.125    0.606     0.304     neutral
                 0.000    0.022    0.000      0.000    0.000      -0.037   0.554     0.072     hate
                 0.000    0.000    ?          0.000    ?          ?        0.406     0.014     enthusiasm
                 0.000    0.003    0.000      0.000    0.000      -0.011   0.455     0.070     love
                 0.000    0.009    0.000      0.000    0.000      -0.018   0.629     0.077     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.548     0.036     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.641     0.018     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.393     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.586     0.014     anger
Weighted Avg.    0.309    0.226    ?          0.309    ?          ?        0.598     0.293

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  4  4  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 31 34  0 17  1  0  0  0  0  0  0  0 |  b = sadness
```

```
   0 19 46   0 23   5   0   1   0   0   0   0   0 |   c = worry
   0   0   2   0   0   0   0   0   1   0   0   0   0 |   d = fun
   1 14 32   0 28   1   0   0   2   0   0   0   0 |   e = neutral
   0 10   9   0   2   0   0   0   0   0   0   0   0 |   f = hate
   0   0   4   0   0   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0   7   3   0   4   0   0   0   0   0   0   0   0 |   h = love
   0   4   4   0   4   0   0   0   0   0   0   0   0 |   i = surprise
   0   4   4   0   2   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   0   0   2   0   0   0   0   0   0   0   0 |   k = boredom
   0   3   0   0   2   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.5 Minimum Frequency = 6

Words will be considered only if they are repeated 6 times at least.

4.5.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         342               34.2    %
Incorrectly Classified Instances       658               65.8    %
Kappa statistic                          0.1176
Mean absolute error                      0.1144
Root mean squared error                  0.2479
Relative absolute error                 93.5231 %
Root relative squared error            100.3276 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.007   0.624     0.034     empty
                 0.390    0.195    0.395      0.390    0.393      0.196    0.636     0.399     sadness
                 0.521    0.410    0.339      0.521    0.411      0.101    0.606     0.386     worry
                 0.000    0.000    ?          0.000    ?          ?        0.418     0.014     fun
                 0.391    0.246    0.329      0.391    0.357      0.138    0.612     0.304     neutral
                 0.019    0.013    0.077      0.019    0.030      0.012    0.592     0.103     hate
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.457     0.014     enthusiasm
                 0.034    0.002    0.333      0.034    0.063      0.099    0.601     0.091     love
                 0.021    0.009    0.100      0.021    0.034      0.024    0.530     0.062     surprise
                 0.034    0.002    0.333      0.034    0.063      0.099    0.608     0.058     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.505     0.007     boredom
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.431     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.622     0.007     anger
Weighted Avg.    0.342    0.225    ?          0.342    ?          ?        0.603     0.295

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  12   0   7   0   0   0   1   0   0   0   0 |   a = empty
   0  96  98   0  43   6   0   1   2   0   0   0   0 |   b = sadness
   0  67 150   0  66   2   2   0   1   0   0   0   0 |   c = worry
   0   3   7   0   5   0   0   0   0   1   0   0   0 |   d = fun
   1  30 104   0  92   3   1   0   3   0   0   1   0 |   e = neutral
   0  15  23   0  15   1   0   0   0   0   0   0   0 |   f = hate
   0   3   6   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0   9   4   0  12   0   0   1   2   1   0   0   0 |   h = love
   0  11  20   0  15   1   0   0   1   0   0   0   0 |   i = surprise
   1   3  11   0  12   0   0   1   0   1   0   0   0 |   j = happiness
   0   2   3   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   3   2   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.5.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         105               30.8824 %
Incorrectly Classified Instances       235               69.1176 %
Kappa statistic                          0.0863
Mean absolute error                      0.1156
Root mean squared error                  0.2519
Relative absolute error                 94.113  %
Root relative squared error            101.3092 %
Total Number of Instances              340
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.637     0.061     empty
                 0.361    0.245    0.323      0.361    0.341      0.112    0.590     0.367     sadness
                 0.500    0.386    0.331      0.500    0.398      0.103    0.615     0.406     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.511     0.013     fun
                 0.333    0.237    0.295      0.333    0.313      0.093    0.586     0.267     neutral
                 0.048    0.013    0.200      0.048    0.077      0.070    0.582     0.149     hate
                 0.000    0.006    0.000      0.000    0.000      -0.008   0.592     0.019     enthusiasm
                 0.000    0.006    0.000      0.000    0.000      -0.016   0.575     0.075     love
                 0.083    0.012    0.200      0.083    0.118      0.109    0.539     0.065     surprise
                 0.000    0.006    0.000      0.000    0.000      -0.013   0.666     0.071     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.607     0.019     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.420     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.580     0.018     anger
Weighted Avg.    0.309    0.223    ?          0.309    ?          ?        0.594     0.283

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  2  6  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 30 32  0 16  1  1  1  2  0  0  0  0 |  b = sadness
  0 20 47  0 25  1  1  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 14 34  0 26  1  0  1  2  0  0  0  0 |  e = neutral
  0  8  7  1  3  1  0  0  0  1  0  0  0 |  f = hate
  0  1  3  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  6  3  0  4  0  0  0  0  1  0  0  0 |  h = love
  0  4  2  0  4  1  0  0  1  0  0  0  0 |  i = surprise
  0  3  4  0  3  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  4  0  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.6 Lovin Stemmer

The first ever published stemming algorithm was: Lovins JB (1968) Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11: 22-31. Julie Beth Lovins' paper was remarkable for the early date at which it was done, and for its seminal influence on later work in this area.[18]

The design of the algorithm was much influenced by the technical vocabulary with which Lovins found herself working (subject term keywords attached to documents in the materials science and engineering field). The subject term list may also have been slightly limiting in that certain common endings are not represented (ements and ents for example, corresponding to the singular forms ement and ent), and also in that the algorithm's treatment of short words, or words with short stems, can be rather destructive.[18]

The Lovins algorithm is noticeably bigger than the Porter algorithm, because of its very extensive endings list. But in one way that is used to advantage: it is faster. It has effectively traded space for time, and with its large suffix set it needs just two major steps to remove a suffix, compared with the eight of the Porter algorithm.[18]

The Lovins stemmer has 294 endings, 29 conditions and 35 transformation rules. Each ending is associated with one of the conditions. In the first step the longest ending is found which satisfies its associated condition, and is removed. In the second step the 35 rules are applied to transform the ending. The second step is done whether or not an ending is removed in the first step.[18]

For example, nationally has the ending ationally, with associated condition, B, 'minimum stem length = 3'. Since removing ationally would leave a stem of length 1 this is rejected. But it also has ending ionally with associated condition A. Condition A is ‘no restriction on stem length’, so ionally is removed, leaving nat.[18]

The transformation rules handle features like letter undoubling (sitting -> sitt -> sit), irregular plurals (matrix and matrices), and English morphological oddities ultimately caused by the behaviour of Latin verbs of the second conjugation (assume / assumption, commit / commission etc). Although they are described as being applied in turn, they can be broken into two stages, rule 1 being done in stage 1, and either zero or one of rules 2 to 35 being done in stage 2. [18]

4.6.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         345               34.5    %
Incorrectly Classified Instances       655               65.5    %
Kappa statistic                          0.1082
Mean absolute error                      0.1049
Root mean squared error                  0.2651
Relative absolute error                 85.7815 %
Root relative squared error            107.3037 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.644     0.059     empty
                 0.390    0.269    0.321      0.390    0.352      0.114    0.603     0.327     sadness
                 0.642    0.487    0.348      0.642    0.451      0.141    0.615     0.386     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.495     0.028     fun
                 0.272    0.131    0.390      0.272    0.321      0.162    0.661     0.345     neutral
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.510     0.067     hate
                 0.000    0.000    ?          0.000    ?          ?        0.463     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.479     0.040     love
                 0.000    0.002    0.000      0.000    0.000      -0.010   0.594     0.065     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.461     0.026     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.582     0.010     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.546     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.784     0.036     anger
Weighted Avg.    0.345    0.238    ?          0.345    ?          ?        0.604     0.284

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  13   0   4   0   0   0   0   0   0   0   0 |   a = empty
   0  96 123   0  27   0   0   0   0   0   0   0   0 |   b = sadness
   0  74 185   1  28   0   0   0   0   0   0   0   0 |   c = worry
   0   3   6   0   6   1   0   0   0   0   0   0   0 |   d = fun
   0  56 114   0  64   0   0   0   1   0   0   0   0 |   e = neutral
   0  21  23   0   9   0   0   0   1   0   0   0   0 |   f = hate
   0   4   6   0   4   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  17   7   0   5   0   0   0   0   0   0   0   0 |   h = love
   0  13  27   0   7   1   0   0   0   0   0   0   0 |   i = surprise
   0   8  17   0   4   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   5   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.6.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         101               29.7059 %
Incorrectly Classified Instances       239               70.2941 %
Kappa statistic                          0.0484
Mean absolute error                      0.108
Root mean squared error                  0.275
Relative absolute error                 87.8918 %
Root relative squared error            110.587  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.707     0.160     empty
                 0.289    0.249    0.273      0.289    0.281      0.039    0.594     0.294     sadness
                 0.681    0.581    0.309      0.681    0.425      0.091    0.603     0.392     worry
                 0.000    0.000    ?          0.000    ?          ?        0.668     0.177     fun
                 0.167    0.118    0.295      0.167    0.213      0.061    0.620     0.310     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.479     0.075     hate
                 0.000    0.000    ?          0.000    ?          ?        0.379     0.016     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.497     0.058     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.720     0.071     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.535     0.034     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.761     0.033     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.478     0.020     relief
                 0.000    0.000    ?          0.000    ?          ?        0.682     0.014     anger
Weighted Avg.    0.297    0.249    ?          0.297    ?          ?        0.596     0.269

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  3  7  0  1  0  0  0  0  0  0  0  0 |  a = empty
  0 24 49  0 10  0  0  0  0  0  0  0  0 |  b = sadness
  0 21 64  0  9  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 18 46  0 13  0  0  0  1  0  0  0  0 |  e = neutral
  0  4 13  0  4  0  0  0  0  0  0  0  0 |  f = hate
  0  0  3  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  5  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  3  8  0  1  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  2  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  2  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.7 Rainbow Stopword

Just a list of stop words, these words has usually no importance in the text and they are repeated quite often, removing them increases the speed of execution and and might increase the performance as we will see later.

4.7.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         321               32.1    %
Incorrectly Classified Instances       679               67.9    %
Kappa statistic                          0.0918
Mean absolute error                      0.1137
Root mean squared error                  0.2521
Relative absolute error                 92.9661 %
Root relative squared error            102.0596 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.596     0.033     empty
                 0.439    0.284    0.335      0.439    0.380      0.143    0.635     0.347     sadness
                 0.434    0.354    0.332      0.434    0.376      0.075    0.573     0.379     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.511     0.019     fun
                 0.357    0.242    0.312      0.357    0.333      0.111    0.631     0.307     neutral
                 0.000    0.006    0.000      0.000    0.000      -0.019   0.597     0.093     hate
                 0.000    0.000    ?          0.000    ?          ?        0.545     0.017     enthusiasm
                 0.034    0.004    0.200      0.034    0.059      0.072    0.586     0.069     love
                 0.063    0.013    0.200      0.063    0.095      0.088    0.610     0.089     surprise
                 0.000    0.005    0.000      0.000    0.000      -0.012   0.586     0.041     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.527     0.011     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.550     0.108     relief
                 0.000    0.000    ?          0.000    ?          ?        0.601     0.005     anger
Weighted Avg.    0.321    0.230    ?          0.321    ?          ?        0.604     0.282

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  10   0   5   1   0   0   1   0   0   0   0 |   a = empty
   0 108  91   0  42   2   0   1   1   1   0   0   0 |   b = sadness
   0  89 125   0  66   1   0   1   4   2   0   0   0 |   c = worry
   0   5   4   0   6   1   0   0   0   0   0   0   0 |   d = fun
   0  59  85   0  84   1   0   2   3   1   0   0   0 |   e = neutral
   0  14  21   1  17   0   0   0   1   0   0   0   0 |   f = hate
   0   3   8   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  12   5   0   9   0   0   1   2   0   0   0   0 |   h = love
   0  12  12   0  20   0   0   0   3   1   0   0   0 |   i = surprise
   0  13  10   0   6   0   0   0   0   0   0   0   0 |   j = happiness
   0   0   1   0   5   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   5   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   1   0   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

4.7.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          99               29.1176 %
Incorrectly Classified Instances       241               70.8824 %
Kappa statistic                          0.0543
Mean absolute error                      0.1147
Root mean squared error                  0.2548
Relative absolute error                 93.368  %
Root relative squared error            102.4627 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.625     0.082     empty
                 0.386    0.272    0.314      0.386    0.346      0.106    0.615     0.350     sadness
                 0.500    0.447    0.299      0.500    0.375      0.047    0.559     0.369     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.703     0.024     fun
                 0.256    0.198    0.278      0.256    0.267      0.060    0.605     0.304     neutral
                 0.000    0.006    0.000      0.000    0.000      -0.020   0.564     0.100     hate
                 0.000    0.000    ?          0.000    ?          ?        0.451     0.051     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.698     0.199     love
                 0.000    0.015    0.000      0.000    0.000      -0.023   0.758     0.121     surprise
                 0.000    0.003    0.000      0.000    0.000      -0.009   0.724     0.079     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.649     0.023     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.635     0.043     relief
                 0.000    0.000    ?          0.000    ?          ?        0.543     0.011     anger
Weighted Avg.    0.291    0.237    ?          0.291    ?          ?        0.605     0.282

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  8  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 32 35  0 14  1  0  0  1  0  0  0  0 |  b = sadness
  0 28 47  0 16  1  0  0  1  1  0  0  0 |  c = worry
  0  0  1  0  1  0  0  0  1  0  0  0  0 |  d = fun
  0 16 41  0 20  0  0  0  1  0  0  0  0 |  e = neutral
  0  6 12  1  2  0  0  0  0  0  0  0  0 |  f = hate
  0  2  2  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  3  0  3  0  0  0  1  0  0  0  0 |  h = love
  0  3  3  0  6  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  2  0  3  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  2  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.8 TF + IDF

4.8.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         204               20.4    %
Incorrectly Classified Instances       796               79.6    %
Kappa statistic                          0.0675
Mean absolute error                      0.123
Root mean squared error                  0.3233
Relative absolute error                100.5082 %
Root relative squared error            130.8636 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.095    0.034    0.057      0.095    0.071      0.048    0.599     0.032     empty
                 0.268    0.172    0.337      0.268    0.299      0.104    0.616     0.319     sadness
                 0.240    0.150    0.392      0.240    0.297      0.106    0.565     0.362     worry
                 0.063    0.026    0.037      0.063    0.047      0.028    0.434     0.025     fun
                 0.170    0.123    0.299      0.170    0.217      0.059    0.585     0.282     neutral
                 0.130    0.086    0.080      0.130    0.099      0.035    0.524     0.065     hate
                 0.071    0.052    0.019      0.071    0.030      0.010    0.497     0.019     enthusiasm
                 0.172    0.063    0.076      0.172    0.105      0.074    0.558     0.045     love
                 0.167    0.102    0.076      0.167    0.105      0.045    0.581     0.065     surprise
                 0.172    0.063    0.076      0.172    0.105      0.074    0.579     0.123     happiness
                 0.000    0.023    0.000      0.000    0.000      -0.012   0.543     0.012     boredom
                 0.000    0.025    0.000      0.000    0.000      -0.017   0.540     0.013     relief
                 0.000    0.007    0.000      0.000    0.000      -0.005   0.751     0.019     anger
Weighted Avg.    0.204    0.130    0.280      0.204    0.229      0.080    0.579     0.262

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  2  3  4  0  4  2  1  1  2  0  1  1  0 |  a = empty
  7 66 45 10 24 20 10 15 24 14  6  5  0 |  b = sadness
 13 52 69  5 36 32 16 17 23 14  4  4  3 |  c = worry
  0  2  2  1  1  3  0  0  3  3  0  1  0 |  d = fun
  8 38 35  7 40 14 11 13 32 15  7 12  3 |  e = neutral
  2  8 10  2  5  7  7  4  5  2  2  0  0 |  f = hate
  0  4  2  0  3  1  1  0  0  3  0  0  0 |  g = enthusiasm
  1 10  0  0  2  1  0  5  4  6  0  0  0 |  h = love
  1  7  4  0 11  5  2  6  8  3  1  0  0 |  i = surprise
  0  4  2  0  3  3  4  4  2  5  0  2  0 |  j = happiness
  1  1  0  0  2  0  0  0  1  0  0  0  1 |  k = boredom
  0  1  2  2  2  0  0  1  1  1  1  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  1  0  0 |  m = anger
```

4.8.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          81               23.8235 %
Incorrectly Classified Instances       259               76.1765 %
Kappa statistic                          0.0839
Mean absolute error                      0.117
Root mean squared error                  0.3123
Relative absolute error                 95.2547 %
Root relative squared error            125.6242 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.091    0.018    0.143      0.091    0.111      0.091    0.689     0.101     empty
                 0.265    0.163    0.344      0.265    0.299      0.112    0.609     0.321     sadness
                 0.351    0.228    0.371      0.351    0.361      0.126    0.581     0.384     worry
                 0.333    0.021    0.125      0.333    0.182      0.193    0.499     0.173     fun
                 0.179    0.156    0.255      0.179    0.211      0.026    0.592     0.279     neutral
                 0.095    0.060    0.095      0.095    0.095      0.036    0.512     0.092     hate
                 0.000    0.048    0.000      0.000    0.000      -0.024   0.406     0.012     enthusiasm
                 0.214    0.058    0.136      0.214    0.167      0.126    0.567     0.077     love
                 0.250    0.088    0.094      0.250    0.136      0.102    0.736     0.087     surprise
                 0.200    0.045    0.118      0.200    0.148      0.120    0.740     0.101     happiness
                 0.000    0.012    0.000      0.000    0.000      -0.010   0.693     0.019     boredom
                 0.000    0.012    0.000      0.000    0.000      -0.013   0.487     0.018     relief
                 0.000    0.003    0.000      0.000    0.000      -0.004   0.630     0.012     anger
Weighted Avg.    0.238    0.151    0.269      0.238    0.248      0.087    0.596     0.269

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  1  1  3  0  2  0  1  0  2  0  1  0  0 |  a = empty
  2 22 19  2 14  2  4  5  7  3  2  1  0 |  b = sadness
  2 17 33  3 13  7  3  2  8  4  1  0  1 |  c = worry
  0  0  1  1  0  0  0  0  1  0  0  0  0 |  d = fun
  0 12 20  1 14  7  2  7 10  3  0  2  0 |  e = neutral
  0  2  6  1  3  2  2  2  1  2  0  0  0 |  f = hate
  0  2  2  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  1  4  2  0  2  0  0  3  0  2  0  0  0 |  h = love
  0  2  1  0  3  2  1  0  3  0  0  0  0 |  i = surprise
  0  1  0  0  2  1  3  1  0  2  0  0  0 |  j = happiness
  1  0  0  0  1  0  0  0  0  0  0  1  0 |  k = boredom
  0  1  1  0  0  0  0  2  0  1  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.9 TF + Lower Case

4.9.1 Cross Validation (10 Folds)

```
=== Summary ===
Correctly Classified Instances         344               34.4    %
Incorrectly Classified Instances       656               65.6    %
Kappa statistic                          0.0996
Mean absolute error                      0.1067
Root mean squared error                  0.2564
Relative absolute error                 87.194  %
Root relative squared error            103.7845 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.639     0.038     empty
                 0.354    0.220    0.344      0.354    0.349      0.132    0.616     0.340     sadness
                 0.753    0.587    0.342      0.753    0.470      0.157    0.636     0.400     worry
                 0.000    0.000    ?          0.000    ?          ?        0.505     0.019     fun
                 0.170    0.094    0.357      0.170    0.231      0.102    0.669     0.340     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.476     0.061     hate
                 0.000    0.000    ?          0.000    ?          ?        0.483     0.016     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.456     0.035     love
                 0.000    0.000    ?          0.000    ?          ?        0.599     0.065     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.428     0.024     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.575     0.010     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.565     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.784     0.023     anger
Weighted Avg.    0.344    0.245    ?          0.344    ?          ?        0.612     0.289

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  14   0   3   0   0   0   0   0   0   0   0 |   a = empty
   0  87 142   0  17   0   0   0   0   0   0   0   0 |   b = sadness
   0  58 217   0  13   0   0   0   0   0   0   0   0 |   c = worry
   0   4   7   0   5   0   0   0   0   0   0   0   0 |   d = fun
   0  55 140   0  40   0   0   0   0   0   0   0   0 |   e = neutral
   0  14  33   0   7   0   0   0   0   0   0   0   0 |   f = hate
   0   1  11   0   2   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  12  12   0   5   0   0   0   0   0   0   0   0 |   h = love
   0   8  29   0  11   0   0   0   0   0   0   0   0 |   i = surprise
   0   8  18   0   3   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   6   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.9.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         107               31.4706 %
Incorrectly Classified Instances       233               68.5294 %
Kappa statistic                          0.0677
Mean absolute error                      0.1083
Root mean squared error                  0.2618
Relative absolute error                 88.1973 %
Root relative squared error            105.3042 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.710     0.150     empty
                 0.313    0.183    0.356      0.313    0.333      0.136    0.606     0.322     sadness
                 0.755    0.679    0.298      0.755    0.428      0.075    0.627     0.412     worry
                 0.000    0.000    ?          0.000    ?          ?        0.650     0.058     fun
                 0.128    0.069    0.357      0.128    0.189      0.091    0.643     0.336     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.471     0.069     hate
                 0.000    0.000    ?          0.000    ?          ?        0.379     0.021     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.474     0.056     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.720     0.079     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.522     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.756     0.036     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.537     0.025     relief
                 0.000    0.000    ?          0.000    ?          ?        0.707     0.018     anger
Weighted Avg.    0.315    0.248    ?          0.315    ?          ?        0.610     0.286

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  8  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 26 55  0  2  0  0  0  0  0  0  0  0 |  b = sadness
  0 18 71  0  5  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 11 56  0 10  0  0  0  1  0  0  0  0 |  e = neutral
  0  3 16  0  2  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  6  7  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  3  7  0  2  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  6  0  0  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  3  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.10 TF + Minimum Frequency = 3

4.10.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         358               35.8    %
Incorrectly Classified Instances       642               64.2    %
Kappa statistic                          0.1309
Mean absolute error                      0.1111
Root mean squared error                  0.2451
Relative absolute error                 90.8186 %
Root relative squared error             99.2057 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.614     0.031     empty
                 0.427    0.212    0.396      0.427    0.411      0.209    0.651     0.404     sadness
                 0.556    0.419    0.349      0.556    0.429      0.125    0.637     0.416     worry
                 0.000    0.000    ?          0.000    ?          ?        0.528     0.019     fun
                 0.396    0.234    0.342      0.396    0.367      0.154    0.648     0.344     neutral
                 0.000    0.004    0.000      0.000    0.000      -0.015   0.559     0.069     hate
                 0.000    0.000    ?          0.000    ?          ?        0.456     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.037     love
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.547     0.057     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.539     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.596     0.009     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.469     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.677     0.008     anger
Weighted Avg.    0.358    0.228    ?          0.358    ?          ?        0.621     0.310

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  14   0   6   0   0   0   0   0   0   0   0 |   a = empty
   0 105  94   0  45   2   0   0   0   0   0   0   0 |   b = sadness
   0  67 160   0  59   1   0   0   1   0   0   0   0 |   c = worry
   0   2   7   0   7   0   0   0   0   0   0   0   0 |   d = fun
   0  37 104   0  93   1   0   0   0   0   0   0   0 |   e = neutral
   0  15  25   0  14   0   0   0   0   0   0   0   0 |   f = hate
   0   2   9   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  13   7   0   9   0   0   0   0   0   0   0   0 |   h = love
   0  11  17   0  20   0   0   0   0   0   0   0   0 |   i = surprise
   0   8  12   0   9   0   0   0   0   0   0   0   0 |   j = happiness
   0   2   1   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   6   0   3   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.10.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         110               32.3529 %
Incorrectly Classified Instances       230               67.6471 %
Kappa statistic                          0.0914
Mean absolute error                      0.1128
Root mean squared error                  0.2499
Relative absolute error                 91.8381 %
Root relative squared error            100.4985 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.658     0.051     empty
                 0.337    0.241    0.311      0.337    0.324      0.094    0.621     0.390     sadness
                 0.543    0.459    0.311      0.543    0.395      0.074    0.626     0.419     worry
                 0.000    0.000    ?          0.000    ?          ?        0.670     0.019     fun
                 0.397    0.206    0.365      0.397    0.380      0.186    0.605     0.310     neutral
                 0.000    0.003    0.000      0.000    0.000      -0.014   0.523     0.065     hate
                 0.000    0.000    ?          0.000    ?          ?        0.388     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.428     0.066     love
                 0.000    0.000    ?          0.000    ?          ?        0.628     0.073     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.536     0.038     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.645     0.018     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.408     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.592     0.014     anger
Weighted Avg.    0.324    0.233    ?          0.324    ?          ?        0.598     0.295

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  4  4  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 28 42  0 12  1  0  0  0  0  0  0  0 |  b = sadness
  0 20 51  0 23  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 10 37  0 31  0  0  0  0  0  0  0  0 |  e = neutral
  0  8 11  0  2  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  3  0  3  0  0  0  0  0  0  0  0 |  h = love
  0  4  4  0  4  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  4  0  2  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  0  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  1  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.11 TF + Minimum Frequency = 6

4.11.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         346               34.6    %
Incorrectly Classified Instances       654               65.4    %
Kappa statistic                          0.1141
Mean absolute error                      0.116
Root mean squared error                  0.2445
Relative absolute error                 94.8156 %
Root relative squared error             98.9827 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.619     0.032     empty
                 0.402    0.192    0.406      0.402    0.404      0.211    0.635     0.398     sadness
                 0.569    0.458    0.335      0.569    0.422      0.101    0.606     0.386     worry
                 0.000    0.000    ?          0.000    ?          ?        0.429     0.014     fun
                 0.345    0.229    0.316      0.345    0.330      0.113    0.610     0.304     neutral
                 0.019    0.001    0.500      0.019    0.036      0.088    0.577     0.100     hate
                 0.000    0.000    ?          0.000    ?          ?        0.456     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.605     0.085     love
                 0.021    0.005    0.167      0.021    0.037      0.043    0.536     0.060     surprise
                 0.000    0.002    0.000      0.000    0.000      -0.008   0.589     0.049     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.509     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.440     0.010     relief
                 0.000    0.000    ?          0.000    ?          ?        0.607     0.006     anger
Weighted Avg.    0.346    0.233    ?          0.346    ?          ?        0.602     0.294

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  12   0   7   0   0   0   1   0   0   0   0 |   a = empty
   0  99 104   0  41   1   0   0   1   0   0   0   0 |   b = sadness
   0  62 164   0  62   0   0   0   0   0   0   0   0 |   c = worry
   0   3   9   0   3   0   0   0   0   1   0   0   0 |   d = fun
   0  29 122   0  81   0   0   0   2   1   0   0   0 |   e = neutral
   0  15  25   0  13   1   0   0   0   0   0   0   0 |   f = hate
   0   3   6   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  10   5   0  13   0   0   0   1   0   0   0   0 |   h = love
   0  11  22   0  14   0   0   0   1   0   0   0   0 |   i = surprise
   0   6  13   0  10   0   0   0   0   0   0   0   0 |   j = happiness
   0   2   3   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   3   3   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.11.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         108               31.7647 %
Incorrectly Classified Instances       232               68.2353 %
Kappa statistic                          0.0868
Mean absolute error                      0.1168
Root mean squared error                  0.2473
Relative absolute error                 95.0736 %
Root relative squared error             99.4621 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.639     0.056     empty
                 0.349    0.241    0.319      0.349    0.333      0.105    0.599     0.374     sadness
                 0.532    0.447    0.313      0.532    0.394      0.076    0.620     0.409     worry
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.012     fun
                 0.346    0.218    0.321      0.346    0.333      0.125    0.586     0.266     neutral
                 0.048    0.003    0.500      0.048    0.087      0.140    0.586     0.144     hate
                 0.000    0.000    ?          0.000    ?          ?        0.582     0.020     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.574     0.069     love
                 0.083    0.006    0.333      0.083    0.133      0.152    0.540     0.065     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.620     0.063     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.613     0.020     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.429     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.583     0.018     anger
Weighted Avg.    0.318    0.233    ?          0.318    ?          ?        0.596     0.284

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  2  6  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 29 40  0 14  0  0  0  0  0  0  0  0 |  b = sadness
  0 20 50  0 23  1  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 13 36  0 27  0  0  0  2  0  0  0  0 |  e = neutral
  0  9  8  0  3  1  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  3  0  3  0  0  0  0  0  0  0  0 |  h = love
  0  3  4  0  4  0  0  0  1  0  0  0  0 |  i = surprise
  0  3  4  0  3  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  1  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.12 TF + Lovin Stemmer

4.12.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         341               34.1    %
Incorrectly Classified Instances       659               65.9    %
Kappa statistic                          0.0953
Mean absolute error                      0.1069
Root mean squared error                  0.2573
Relative absolute error                 87.3836 %
Root relative squared error            104.1608 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.634     0.044     empty
                 0.333    0.214    0.337      0.333    0.335      0.120    0.604     0.335     sadness
                 0.743    0.603    0.333      0.743    0.460      0.133    0.621     0.386     worry
                 0.000    0.000    ?          0.000    ?          ?        0.488     0.020     fun
                 0.191    0.089    0.398      0.191    0.259      0.137    0.667     0.354     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.064     hate
                 0.000    0.000    ?          0.000    ?          ?        0.467     0.014     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.470     0.037     love
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.592     0.064     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.443     0.025     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.585     0.010     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.549     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.783     0.031     anger
Weighted Avg.    0.341    0.247    ?          0.341    ?          ?        0.605     0.287

=== Confusion Matrix ===
   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  14   0   3   0   0   0   0   0   0   0   0 |   a = empty
   0  82 142   0  22   0   0   0   0   0   0   0   0 |   b = sadness
   0  57 214   0  17   0   0   0   0   0   0   0   0 |   c = worry
   0   4   9   0   3   0   0   0   0   0   0   0   0 |   d = fun
   0  46 143   0  45   0   0   0   1   0   0   0   0 |   e = neutral
   0  15  34   0   5   0   0   0   0   0   0   0   0 |   f = hate
   0   1  12   0   1   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  14  10   0   5   0   0   0   0   0   0   0   0 |   h = love
   0   9  32   0   7   0   0   0   0   0   0   0   0 |   i = surprise
   0   9  20   0   0   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   7   0   3   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

4.12.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          94               27.6471 %
Incorrectly Classified Instances       246               72.3529 %
Kappa statistic                          0.0154
Mean absolute error                      0.1091
Root mean squared error                  0.2648
Relative absolute error                 88.8153 %
Root relative squared error            106.4949 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.700     0.155     empty
                 0.205    0.214    0.236      0.205    0.219      -0.010   0.588     0.294     sadness
                 0.723    0.699    0.283      0.723    0.407      0.024    0.615     0.403     worry
                 0.000    0.000    ?          0.000    ?          ?        0.656     0.177     fun
                 0.115    0.069    0.333      0.115    0.171      0.073    0.635     0.322     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.465     0.072     hate
                 0.000    0.000    ?          0.000    ?          ?        0.388     0.017     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.491     0.063     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.711     0.070     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.526     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.752     0.033     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.482     0.020     relief
                 0.000    0.000    ?          0.000    ?          ?        0.688     0.014     anger
Weighted Avg.    0.276    0.261    ?          0.276    ?          ?        0.600     0.275

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  2  7  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 17 62  0  4  0  0  0  0  0  0  0  0 |  b = sadness
  0 20 68  0  6  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 15 53  0  9  0  0  0  1  0  0  0  0 |  e = neutral
  0  4 16  0  1  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  6  7  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  2  9  0  1  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  4  0  0  0  0  0  0  0  0  0  0 |  l = relief
  0  0  2  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.13 TF + Rainbow Stopword

4.13.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         319               31.9    %
Incorrectly Classified Instances       681               68.1    %
Kappa statistic                          0.0758
Mean absolute error                      0.1155
Root mean squared error                  0.2462
Relative absolute error                 94.3717 %
Root relative squared error             99.6612 %
Total Number of Instances             1000
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.587     0.032     empty
                 0.427    0.268    0.342      0.427    0.380      0.148    0.636     0.351     sadness
                 0.524    0.454    0.319      0.524    0.396      0.064    0.579     0.379     worry
                 0.000    0.000    ?          0.000    ?          ?        0.510     0.019     fun
                 0.268    0.201    0.290      0.268    0.279      0.069    0.625     0.308     neutral
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.598     0.101     hate
                 0.000    0.000    ?          0.000    ?          ?        0.545     0.017     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.601     0.076     love
                 0.000    0.000    ?          0.000    ?          ?        0.612     0.089     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.570     0.038     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.530     0.011     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.542     0.064     relief
                 0.000    0.000    ?          0.000    ?          ?        0.567     0.005     anger
Weighted Avg.    0.319    0.244    ?          0.319    ?          ?        0.605     0.283

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   4  12   0   5   0   0   0   0   0   0   0   0 |   a = empty
   0 105 108   0  32   1   0   0   0   0   0   0   0 |   b = sadness
   0  80 151   0  57   0   0   0   0   0   0   0   0 |   c = worry
   0   4   6   0   5   1   0   0   0   0   0   0   0 |   d = fun
   0  54 118   0  63   0   0   0   0   0   0   0   0 |   e = neutral
   0  15  24   0  15   0   0   0   0   0   0   0   0 |   f = hate
   0   3   8   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  13   8   0   8   0   0   0   0   0   0   0   0 |   h = love
   0  13  18   0  17   0   0   0   0   0   0   0   0 |   i = surprise
   0  13  11   0   5   0   0   0   0   0   0   0   0 |   j = happiness
   0   0   3   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   5   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   1   2   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.13.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          98               28.8235 %
Incorrectly Classified Instances       242               71.1765 %
Kappa statistic                          0.0381
Mean absolute error                      0.1166
Root mean squared error                  0.2488
Relative absolute error                 94.9467 %
Root relative squared error            100.0683 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.621     0.078     empty
                 0.349    0.249    0.312      0.349    0.330      0.097    0.615     0.359     sadness
                 0.596    0.553    0.292      0.596    0.392      0.039    0.567     0.365     worry
                 0.000    0.000    ?          0.000    ?          ?        0.710     0.024     fun
                 0.167    0.160    0.236      0.167    0.195      0.007    0.595     0.298     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.560     0.089     hate
                 0.000    0.000    ?          0.000    ?          ?        0.462     0.040     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.701     0.176     love
                 0.000    0.000    ?          0.000    ?          ?        0.762     0.132     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.710     0.077     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.642     0.023     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.626     0.045     relief
                 0.000    0.000    ?          0.000    ?          ?        0.546     0.011     anger
Weighted Avg.    0.288    0.250    ?          0.288    ?          ?        0.605     0.281

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  8  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 29 43  0 11  0  0  0  0  0  0  0  0 |  b = sadness
  0 24 56  0 14  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0 14 51  0 13  0  0  0  0  0  0  0  0 |  e = neutral
  0  7 12  0  2  0  0  0  0  0  0  0  0 |  f = hate
  0  2  2  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  4  0  3  0  0  0  0  0  0  0  0 |  h = love
  0  3  5  0  4  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  2  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.14 Lower Case + Minimum Frequency = 3

4.14.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         362               36.2    %
Incorrectly Classified Instances       638               63.8    %
Kappa statistic                          0.1461
Mean absolute error                      0.1077
Root mean squared error                  0.2505
Relative absolute error                 88.0723 %
Root relative squared error            101.4142 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.007   0.600     0.030     empty
                 0.439    0.232    0.382      0.439    0.408      0.198    0.661     0.414     sadness
                 0.535    0.354    0.379      0.535    0.444      0.167    0.643     0.428     worry
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.504     0.026     fun
                 0.409    0.234    0.349      0.409    0.376      0.166    0.660     0.336     neutral
                 0.056    0.017    0.158      0.056    0.082      0.064    0.579     0.086     hate
                 0.000    0.000    ?          0.000    ?          ?        0.511     0.018     enthusiasm
                 0.034    0.003    0.250      0.034    0.061      0.083    0.543     0.055     love
                 0.000    0.009    0.000      0.000    0.000      -0.021   0.578     0.062     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.537     0.035     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.008     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.481     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.662     0.009     anger
Weighted Avg.    0.362    0.216    ?          0.362    ?          ?        0.631     0.316

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  14   0   6   0   0   0   0   0   0   0   0 |   a = empty
   1 108  80   0  48   6   0   1   2   0   0   0   0 |   b = sadness
   1  64 154   0  61   6   0   0   2   0   0   0   0 |   c = worry
   0   4   5   0   6   0   0   0   1   0   0   0   0 |   d = fun
   0  45  88   0  96   3   0   1   2   0   0   0   0 |   e = neutral
   0  16  22   2  11   3   0   0   0   0   0   0   0 |   f = hate
   0   2   7   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  18   3   0   6   0   0   1   1   0   0   0   0 |   h = love
   0  13  15   0  18   1   0   1   0   0   0   0   0 |   i = surprise
   0   8  10   0  10   0   0   0   1   0   0   0   0 |   j = happiness
   0   2   3   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   2   0   7   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.14.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         112               32.9412 %
Incorrectly Classified Instances       228               67.0588 %
Kappa statistic                          0.1103
Mean absolute error                      0.1107
Root mean squared error                  0.2581
Relative absolute error                 90.1282 %
Root relative squared error            103.7936 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.675     0.057     empty
                 0.386    0.230    0.352      0.386    0.368      0.151    0.627     0.364     sadness
                 0.511    0.407    0.324      0.511    0.397      0.094    0.608     0.426     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.569     0.016     fun
                 0.410    0.210    0.368      0.410    0.388      0.193    0.628     0.316     neutral
                 0.000    0.022    0.000      0.000    0.000      -0.037   0.608     0.086     hate
                 0.000    0.000    ?          0.000    ?          ?        0.408     0.015     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.484     0.079     love
                 0.000    0.012    0.000      0.000    0.000      -0.021   0.643     0.075     surprise
                 0.000    0.006    0.000      0.000    0.000      -0.013   0.556     0.042     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.641     0.040     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.406     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.584     0.012     anger
Weighted Avg.    0.329    0.219    ?          0.329    ?          ?        0.609     0.294

=== Confusion Matrix ===
  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  7  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 32 35  0 11  2  0  0  1  2  0  0  0 |  b = sadness
  0 19 48  0 23  4  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  0  0  0  0  1  0  0  0  0 |  d = fun
  0 13 30  0 32  1  0  0  2  0  0  0  0 |  e = neutral
  0  8  9  1  3  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  4  0  2  0  0  0  0  0  0  0  0 |  h = love
  0  3  2  0  7  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.15 TF + Lower Case + Minimum Frequency = 3

4.15.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         374               37.4    %
Incorrectly Classified Instances       626               62.6    %
Kappa statistic                          0.1531
Mean absolute error                      0.1103
Root mean squared error                  0.2449
Relative absolute error                 90.1928 %
Root relative squared error             99.1119 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.029     empty
                 0.447    0.223    0.396      0.447    0.420      0.216    0.662     0.417     sadness
                 0.580    0.403    0.368      0.580    0.450      0.161    0.645     0.427     worry
                 0.000    0.000    ?          0.000    ?          ?        0.508     0.027     fun
                 0.413    0.214    0.372      0.413    0.391      0.192    0.656     0.331     neutral
                 0.000    0.005    0.000      0.000    0.000      -0.017   0.565     0.083     hate
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.018     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.523     0.055     love
                 0.000    0.002    0.000      0.000    0.000      -0.010   0.565     0.059     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.525     0.033     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.008     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.496     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.661     0.008     anger
Weighted Avg.    0.374    0.222    ?          0.374    ?          ?        0.629     0.315

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  15   0   5   0   0   0   0   0   0   0   0 |   a = empty
   0 110  90   0  41   4   0   0   1   0   0   0   0 |   b = sadness
   0  62 167   0  57   1   0   0   1   0   0   0   0 |   c = worry
   0   4   5   0   7   0   0   0   0   0   0   0   0 |   d = fun
   0  43  95   0  97   0   0   0   0   0   0   0   0 |   e = neutral
   0  17  26   0  11   0   0   0   0   0   0   0   0 |   f = hate
   0   1   8   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  19   4   0   6   0   0   0   0   0   0   0   0 |   h = love
   0  11  20   0  17   0   0   0   0   0   0   0   0 |   i = surprise
   0   7  14   0   8   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   3   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.15.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         118               34.7059 %
Incorrectly Classified Instances       222               65.2941 %
Kappa statistic                          0.1266
Mean absolute error                      0.1122
Root mean squared error                  0.2498
Relative absolute error                 91.3624 %
Root relative squared error            100.4543 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.666     0.054     empty
                 0.398    0.218    0.371      0.398    0.384      0.176    0.633     0.371     sadness
                 0.564    0.427    0.335      0.564    0.421      0.123    0.617     0.427     worry
                 0.000    0.000    ?          0.000    ?          ?        0.585     0.017     fun
                 0.410    0.214    0.364      0.410    0.386      0.189    0.628     0.317     neutral
                 0.000    0.006    0.000      0.000    0.000      -0.020   0.595     0.080     hate
                 0.000    0.000    ?          0.000    ?          ?        0.398     0.015     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.463     0.077     love
                 0.000    0.009    0.000      0.000    0.000      -0.018   0.635     0.066     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.547     0.042     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.639     0.043     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.418     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.012     anger
Weighted Avg.    0.347    0.221    ?          0.347    ?          ?        0.610     0.296

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  6  0  4  0  0  0  0  0  0  0  0 |  a = empty
  0 33 35  0 13  1  0  0  1  0  0  0  0 |  b = sadness
  0 19 53  0 21  1  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  0  0  0  0  1  0  0  0  0 |  d = fun
  0 12 33  0 32  0  0  0  1  0  0  0  0 |  e = neutral
  0  7 10  0  4  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  5  0  2  0  0  0  0  0  0  0  0 |  h = love
  0  3  2  0  7  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.16 Lower Case + Lovin Stemmer + Minimum Frequency = 3

4.16.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         371               37.1    %
Incorrectly Classified Instances       629               62.9    %
Kappa statistic                          0.1593
Mean absolute error                      0.1073
Root mean squared error                  0.2505
Relative absolute error                 87.7238 %
Root relative squared error            101.3985 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.007   0.638     0.035     empty
                 0.463    0.241    0.385      0.463    0.421      0.209    0.652     0.404     sadness
                 0.531    0.340    0.387      0.531    0.448      0.177    0.639     0.422     worry
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.504     0.025     fun
                 0.417    0.225    0.363      0.417    0.388      0.184    0.665     0.347     neutral
                 0.056    0.020    0.136      0.056    0.079      0.055    0.632     0.108     hate
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.477     0.016     enthusiasm
                 0.034    0.003    0.250      0.034    0.061      0.083    0.539     0.052     love
                 0.021    0.007    0.125      0.021    0.036      0.032    0.606     0.081     surprise
                 0.034    0.000    1.000      0.034    0.067      0.183    0.531     0.067     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.451     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.504     0.020     relief
                 0.000    0.000    ?          0.000    ?          ?        0.681     0.010     anger
Weighted Avg.    0.371    0.212    ?          0.371    ?          ?        0.633     0.317

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  12   0   8   0   0   0   0   0   0   0   0 |   a = empty
   0 114  71   0  52   6   1   1   1   0   0   0   0 |   b = sadness
   1  71 153   0  51  10   0   0   2   0   0   0   0 |   c = worry
   0   3   5   0   6   1   0   0   1   0   0   0   0 |   d = fun
   1  47  83   0  98   2   0   2   2   0   0   0   0 |   e = neutral
   0  17  21   1  12   3   0   0   0   0   0   0   0 |   f = hate
   0   3   6   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  14   5   0   9   0   0   1   0   0   0   0   0 |   h = love
   0  13  18   0  16   0   0   0   1   0   0   0   0 |   i = surprise
   0   7  12   0   8   0   0   0   1   1   0   0   0 |   j = happiness
   0   2   3   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   4   3   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.16.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         115               33.8235 %
Incorrectly Classified Instances       225               66.1765 %
Kappa statistic                          0.125
Mean absolute error                      0.1107
Root mean squared error                  0.2591
Relative absolute error                 90.0962 %
Root relative squared error            104.2176 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.061     empty
                 0.410    0.230    0.366      0.410    0.386      0.174    0.622     0.347     sadness
                 0.511    0.370    0.345      0.511    0.412      0.128    0.606     0.406     worry
                 0.000    0.000    ?          0.000    ?          ?        0.577     0.018     fun
                 0.397    0.229    0.341      0.397    0.367      0.160    0.628     0.323     neutral
                 0.048    0.028    0.100      0.048    0.065      0.028    0.645     0.096     hate
                 0.000    0.000    ?          0.000    ?          ?        0.384     0.013     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.520     0.064     love
                 0.083    0.015    0.167      0.083    0.111      0.095    0.675     0.085     surprise
                 0.000    0.003    0.000      0.000    0.000      -0.009   0.541     0.036     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.609     0.036     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.358     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.538     0.009     anger
Weighted Avg.    0.338    0.213    ?          0.338    ?          ?        0.610     0.287

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  2  7  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 34 29  0 14  4  0  0  1  1  0  0  0 |  b = sadness
  0 18 48  0 24  3  0  0  1  0  0  0  0 |  c = worry
  0  0  1  0  1  0  0  0  1  0  0  0  0 |  d = fun
  0 15 29  0 31  1  0  0  2  0  0  0  0 |  e = neutral
  0  7  9  0  4  1  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  6  3  0  4  1  0  0  0  0  0  0  0 |  h = love
  0  3  3  0  5  0  0  0  1  0  0  0  0 |  i = surprise
  0  4  4  0  2  0  0  0  0  0  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  2  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.17 TF + Lower Case + Minimum Frequency = 3 + Rainbow Stopword

4.17.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         354               35.4    %
Incorrectly Classified Instances       646               64.6    %
Kappa statistic                          0.1228
Mean absolute error                      0.1162
Root mean squared error                  0.2419
Relative absolute error                 95.0034 %
Root relative squared error             97.9298 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.511     0.022     empty
                 0.451    0.207    0.416      0.451    0.433      0.238    0.668     0.429     sadness
                 0.569    0.459    0.334      0.569    0.421      0.100    0.636     0.425     worry
                 0.000    0.000    ?          0.000    ?          ?        0.513     0.016     fun
                 0.336    0.210    0.329      0.336    0.333      0.125    0.640     0.320     neutral
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.655     0.115     hate
                 0.000    0.000    ?          0.000    ?          ?        0.606     0.023     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.632     0.116     love
                 0.000    0.000    ?          0.000    ?          ?        0.567     0.056     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.575     0.045     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.469     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.545     0.018     relief
                 0.000    0.000    ?          0.000    ?          ?        0.357     0.003     anger
Weighted Avg.    0.354    0.233    ?          0.354    ?          ?        0.633     0.318

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   0  18   0   3   0   0   0   0   0   0   0   0 |   a = empty
   0 111  94   0  40   1   0   0   0   0   0   0   0 |   b = sadness
   0  61 164   0  62   1   0   0   0   0   0   0   0 |   c = worry
   0   2   6   0   8   0   0   0   0   0   0   0   0 |   d = fun
   0  38 118   0  79   0   0   0   0   0   0   0   0 |   e = neutral
   0  16  28   0  10   0   0   0   0   0   0   0   0 |   f = hate
   0   3   9   0   2   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  12   9   0   8   0   0   0   0   0   0   0   0 |   h = love
   0   9  25   0  14   0   0   0   0   0   0   0   0 |   i = surprise
   0  12  10   0   7   0   0   0   0   0   0   0   0 |   j = happiness
   0   0   3   0   3   0   0   0   0   0   0   0   0 |   k = boredom
   0   3   4   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.17.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         113               33.2353 %
Incorrectly Classified Instances       227               66.7647 %
Kappa statistic                          0.1012
Mean absolute error                      0.1181
Root mean squared error                  0.2462
Relative absolute error                 96.1585 %
Root relative squared error             99.0221 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.538     0.043     empty
                 0.434    0.214    0.396      0.434    0.414      0.213    0.642     0.405     sadness
                 0.585    0.492    0.313      0.585    0.407      0.083    0.576     0.385     worry
                 0.000    0.000    ?          0.000    ?          ?        0.605     0.019     fun
                 0.282    0.191    0.306      0.282    0.293      0.094    0.592     0.295     neutral
                 0.000    0.003    0.000      0.000    0.000      -0.014   0.603     0.116     hate
                 0.000    0.000    ?          0.000    ?          ?        0.597     0.030     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.680     0.143     love
                 0.000    0.000    ?          0.000    ?          ?        0.658     0.093     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.486     0.032     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.580     0.016     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.431     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.369     0.007     anger
Weighted Avg.    0.332    0.232    ?          0.332    ?          ?        0.598     0.293

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  8  0  2  0  0  0  0  0  0  0  0 |  a = empty
  0 36 35  0 11  1  0  0  0  0  0  0  0 |  b = sadness
  0 18 55  0 21  0  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0  9 47  0 22  0  0  0  0  0  0  0  0 |  e = neutral
  0  9  8  0  4  0  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  8  4  0  2  0  0  0  0  0  0  0  0 |  h = love
  0  2  4  0  6  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  3  0  0  0  0  0  0  0  0  0  0 |  l = relief
  0  1  1  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.18 TF + Lower Case + Minimum Frequency = 3 + Lovin Stemmer + Rainbow Stopword

4.18.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         359               35.9    %
Incorrectly Classified Instances       641               64.1    %
Kappa statistic                          0.1316
Mean absolute error                      0.1145
Root mean squared error                  0.2423
Relative absolute error                 93.5509 %
Root relative squared error             98.071  %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.588     0.027     empty
                 0.488    0.232    0.407      0.488    0.444      0.241    0.670     0.421     sadness
                 0.545    0.424    0.342      0.545    0.420      0.110    0.627     0.411     worry
                 0.000    0.000    ?          0.000    ?          ?        0.580     0.026     fun
                 0.349    0.209    0.339      0.349    0.344      0.138    0.654     0.333     neutral
                 0.000    0.004    0.000      0.000    0.000      -0.015   0.648     0.127     hate
                 0.000    0.000    ?          0.000    ?          ?        0.542     0.019     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.647     0.088     love
                 0.000    0.000    ?          0.000    ?          ?        0.597     0.109     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.655     0.075     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.464     0.042     relief
                 0.000    0.000    ?          0.000    ?          ?        0.373     0.003     anger
Weighted Avg.    0.359    0.229    ?          0.359    ?          ?        0.639     0.319

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  17   0   3   0   0   0   0   0   0   0   0 |   a = empty
   0 120  87   0  38   1   0   0   0   0   0   0   0 |   b = sadness
   0  66 157   0  63   2   0   0   0   0   0   0   0 |   c = worry
   0   3   6   0   7   0   0   0   0   0   0   0   0 |   d = fun
   0  50 102   0  82   1   0   0   0   0   0   0   0 |   e = neutral
   0  16  29   0   9   0   0   0   0   0   0   0   0 |   f = hate
   0   3   8   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  13   5   0  11   0   0   0   0   0   0   0   0 |   h = love
   0  10  25   0  13   0   0   0   0   0   0   0   0 |   i = surprise
   0  10  12   0   7   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   4   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.18.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         116               34.1176 %
Incorrectly Classified Instances       224               65.8824 %
Kappa statistic                          0.1151
Mean absolute error                      0.1166
Root mean squared error                  0.2467
Relative absolute error                 94.9158 %
Root relative squared error             99.2275 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.516     0.043     empty
                 0.470    0.226    0.402      0.470    0.433      0.232    0.645     0.390     sadness
                 0.574    0.451    0.327      0.574    0.417      0.110    0.591     0.383     worry
                 0.000    0.000    ?          0.000    ?          ?        0.662     0.024     fun
                 0.295    0.202    0.303      0.295    0.299      0.093    0.609     0.306     neutral
                 0.000    0.006    0.000      0.000    0.000      -0.020   0.586     0.101     hate
                 0.000    0.000    ?          0.000    ?          ?        0.644     0.027     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.746     0.179     love
                 0.000    0.000    ?          0.000    ?          ?        0.628     0.124     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.645     0.065     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.646     0.017     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.260     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.271     0.006     anger
Weighted Avg.    0.341    0.227    ?          0.341    ?          ?        0.610     0.294
```

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  7  0  4  0  0  0  0  0  0  0  0 |  a = empty
  0 39 32  0 11  1  0  0  0  0  0  0  0 |  b = sadness
  0 17 54  0 22  1  0  0  0  0  0  0  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  0 15 40  0 23  0  0  0  0  0  0  0  0 |  e = neutral
  0  8 11  0  2  0  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  3  0  4  0  0  0  0  0  0  0  0 |  h = love
  0  2  6  0  4  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  1  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  2  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.19 TF + Lower Case + Minimum Frequency = 6 + Lovin Stemmer + Rainbow Stopword

4.19.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         360               36      %
Incorrectly Classified Instances       640               64      %
Kappa statistic                          0.128
Mean absolute error                      0.1175
Root mean squared error                  0.2426
Relative absolute error                 96.0054 %
Root relative squared error             98.1911 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.597     0.026     empty
                 0.402    0.195    0.402      0.402    0.402      0.207    0.644     0.417     sadness
                 0.649    0.489    0.350      0.649    0.454      0.146    0.634     0.408     worry
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.021     fun
                 0.315    0.190    0.338      0.315    0.326      0.129    0.614     0.299     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.668     0.165     hate
                 0.000    0.000    ?          0.000    ?          ?        0.553     0.018     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.681     0.134     love
                 0.000    0.000    ?          0.000    ?          ?        0.566     0.059     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.696     0.078     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.453     0.006     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.407     0.010     relief
                 0.000    0.000    ?          0.000    ?          ?        0.218     0.002     anger
Weighted Avg.    0.360    0.233    ?          0.360    ?          ?        0.625     0.309

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   2  16   0   3   0   0   0   0   0   0   0   0 |   a = empty
   0  99 106   0  41   0   0   0   0   0   0   0   0 |   b = sadness
   0  60 187   0  41   0   0   0   0   0   0   0   0 |   c = worry
   0   4   9   0   3   0   0   0   0   0   0   0   0 |   d = fun
   0  35 126   0  74   0   0   0   0   0   0   0   0 |   e = neutral
   0  16  29   0   9   0   0   0   0   0   0   0   0 |   f = hate
   0   3   8   0   3   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  11   6   0  12   0   0   0   0   0   0   0   0 |   h = love
   0   4  26   0  18   0   0   0   0   0   0   0   0 |   i = surprise
   0   9  10   0  10   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   5   0   4   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

4.19.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         113               33.2353 %
Incorrectly Classified Instances       227               66.7647 %
Kappa statistic                          0.0996
Mean absolute error                      0.1189
```

```
Root mean squared error                  0.2456
Relative absolute error                 96.8135 %
Root relative squared error             98.7695 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.510     0.039     empty
                 0.422    0.195    0.412      0.422    0.417      0.225    0.623     0.399     sadness
                 0.638    0.520    0.319      0.638    0.426      0.106    0.615     0.395     worry
                 0.000    0.000    ?          0.000    ?          ?        0.649     0.019     fun
                 0.231    0.183    0.273      0.231    0.250      0.051    0.581     0.269     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.616     0.109     hate
                 0.000    0.000    ?          0.000    ?          ?        0.709     0.027     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.849     0.313     love
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.553     0.055     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.740     0.080     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.557     0.014     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.289     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.168     0.005     anger
Weighted Avg.    0.332    0.233    ?          0.332    ?          ?        0.610     0.294

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  8  0  3  0  0  0  0  0  0  0  0 |  a = empty
  0 35 37  0 10  0  0  0  1  0  0  0  0 |  b = sadness
  0 15 60  0 19  0  0  0  0  0  0  0  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  0 12 48  0 18  0  0  0  0  0  0  0  0 |  e = neutral
  0  8 11  0  2  0  0  0  0  0  0  0  0 |  f = hate
  0  1  3  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  5  4  0  5  0  0  0  0  0  0  0  0 |  h = love
  0  1  8  0  3  0  0  0  0  0  0  0  0 |  i = surprise
  0  4  3  0  3  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  4  1  0  0  0  0  0  0  0  0  0  0 |  l = relief
  0  0  2  0  0  0  0  0  0  0  0  0  0 |  m = anger
```

4.20 TF + IDF + Lower Case + Minimum Frequency = 3 + Lovin Stemmer + Rainbow Stopword

4.20.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         270               27      %
Incorrectly Classified Instances       730               73      %
Kappa statistic                          0.1079
Mean absolute error                      0.1134
Root mean squared error                  0.2905
Relative absolute error                 92.678  %
Root relative squared error            117.5784 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.037    0.000      0.000    0.000      -0.028   0.659     0.034     empty
                 0.341    0.162    0.408      0.341    0.372      0.191    0.645     0.393     sadness
                 0.337    0.237    0.365      0.337    0.350      0.102    0.592     0.399     worry
                 0.000    0.024    0.000      0.000    0.000      -0.020   0.593     0.025     fun
                 0.272    0.157    0.348      0.272    0.305      0.126    0.669     0.348     neutral
                 0.167    0.063    0.130      0.167    0.146      0.092    0.625     0.103     hate
                 0.000    0.023    0.000      0.000    0.000      -0.018   0.511     0.019     enthusiasm
                 0.138    0.048    0.078      0.138    0.100      0.068    0.608     0.052     love
                 0.167    0.059    0.125      0.167    0.143      0.094    0.648     0.098     surprise
                 0.103    0.036    0.079      0.103    0.090      0.059    0.598     0.131     happiness
                 0.000    0.011    0.000      0.000    0.000      -0.008   0.506     0.007     boredom
                 0.091    0.020    0.048      0.091    0.063      0.051    0.512     0.026     relief
                 0.000    0.007    0.000      0.000    0.000      -0.005   0.543     0.005     anger
Weighted Avg.    0.270    0.156    0.305      0.270    0.285      0.119    0.627     0.311

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1 10  0  4  2  0  0  2  1  0  1  0 |  a = empty
  7 84 53  7 36 20  2 10 13  6  3  4  1 |  b = sadness
 10 48 97  5 37 24  9  9 20 11  6  8  4 |  c = worry
```

```
  0  3  2  0  5  2  1  0  3  0  0  0  0 |  d = fun
 12 39 56  7 64  9  5 16 12  9  1  4  1 |  e = neutral
  6  8 13  2  7  9  3  1  2  1  1  1  0 |  f = hate
  0  1  5  0  2  1  0  3  0  2  0  0  0 |  g = enthusiasm
  0  6  4  0  8  1  1  4  2  2  0  1  0 |  h = love
  1  7 12  1 10  1  0  5  8  2  0  0  1 |  i = surprise
  0  5  9  1  3  0  2  3  2  3  0  1  0 |  j = happiness
  0  1  2  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  1  1  4  0  0  0  0  1  0  1  0 |  l = relief
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.20.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          87               25.5882 %
Incorrectly Classified Instances       253               74.4118 %
Kappa statistic                          0.1037
Mean absolute error                      0.1178
Root mean squared error                  0.3
Relative absolute error                 95.9036 %
Root relative squared error            120.6539 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.036    0.000      0.000    0.000      -0.035   0.629     0.060     empty
                 0.349    0.148    0.433      0.349    0.387      0.218    0.612     0.352     sadness
                 0.309    0.260    0.312      0.309    0.310      0.049    0.540     0.365     worry
                 0.000    0.059    0.000      0.000    0.000      -0.024   0.550     0.013     fun
                 0.244    0.126    0.365      0.244    0.292      0.137    0.646     0.331     neutral
                 0.143    0.044    0.176      0.143    0.158      0.109    0.523     0.098     hate
                 0.250    0.036    0.077      0.250    0.118      0.120    0.598     0.063     enthusiasm
                 0.214    0.043    0.176      0.214    0.194      0.156    0.637     0.110     love
                 0.083    0.055    0.053      0.083    0.065      0.023    0.759     0.088     surprise
                 0.200    0.033    0.154      0.200    0.174      0.147    0.668     0.092     happiness
                 0.000    0.012    0.000      0.000    0.000      -0.010   0.744     0.022     boredom
                 0.000    0.033    0.000      0.000    0.000      -0.022   0.374     0.013     relief
                 0.000    0.006    0.000      0.000    0.000      -0.006   0.374     0.007     anger
Weighted Avg.    0.256    0.147    0.301      0.256    0.274      0.116    0.598     0.282

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  5  2  2  0  1  0  1  0  0  0  0 |  a = empty
  2 29 16  7  5  5  2  5  3  3  2  4  0 |  b = sadness
  5 14 29  5 13  5  3  3  7  4  0  5  1 |  c = worry
  1  0  1  0  1  0  0  0  0  0  0  0  0 |  d = fun
  2 11 24  1 19  4  3  4  6  2  0  1  1 |  e = neutral
  1  4  6  2  1  3  1  0  1  1  1  0  0 |  f = hate
  0  1  1  0  1  0  1  0  0  0  0  0  0 |  g = enthusiasm
  1  2  2  0  2  0  1  3  0  1  1  1  0 |  h = love
  0  0  3  3  3  0  0  2  1  0  0  0  0 |  i = surprise
  0  2  4  0  1  0  1  0  0  2  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  0  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

4.21 Conclusion

Correctly Classified Instances in Multinomial Naive Bayes				
Evaluation Process	Cross Validation 10 Folds	Percentage Split 66%	AVG Algorithms	Rank Algorithms
Default	34.9000%	32.3529%	33.6265%	9
IDF	16.8000%	20.8824%	18.8412%	21
TF	34.2000%	30.2941%	32.2471%	14
LowerCase	34.8000%	32.6471%	33.7236%	8
MinFreq=3	36.1000%	30.8824%	33.4912%	10
MinFreq=6	34.2000%	30.8824%	32.5412%	13
LovinsStemmer	34.5000%	29.7059%	32.1030%	15
RainBowStopWord	32.1000%	29.1176%	30.6088%	17
TF+IDF	20.4000%	23.8235%	22.1118%	20
TF+LowerCase	34.4000%	31.4706%	32.9353%	12
TF+MinFreq=3	35.8000%	32.3529%	34.0765%	7
TF+MinFreq=6	34.6000%	31.7647%	33.1824%	11
TF+LovingStemmer	34.1000%	27.6471%	30.8736%	16
TF+RainBowStopWord	31.9000%	28.8235%	30.3618%	18
LoweerCase+MinFreq=3	36.2000%	32.9412%	34.5706%	5
TF+LowerCase+MinFreq=3	37.4000%	34.7059%	36.0530%	1
LowerCase+LovinStemmer+MinFreq=3	37.1000%	33.8235%	35.4618%	2
TF+LowerCase+MinFreq=3+RainBowStopWord	35.4000%	33.2353%	34.3177%	6
TF+LowerCase+MinFreq=3+LovinStemmer+RainBowStopWord	35.9000%	34.1176%	35.0088%	3
TF+LowerCase+MinFreq=6+LovinStemmer+RainBowStopWord	36.0000%	33.2353%	34.6177%	4
TF+IDF+LowerCase+MinFreq=3+RainBowStopWord+LovinStemmer	27.0000%	25.5882%	26.2941%	19
AVG for Evaluation process	33.04%	30.49%	31.76%	

Looking at single options only we see that only Lower Case option was able to beat the original dataset which is weird as I thought that options like stemming and removing stop words will help but they didn't at least on their own.

When combining different options we see a slight improvement. Th best result was obtained by combining Term Frequency representation with lower case option and Minimum word frequency set to 3.

The Second best result was obtained by combining Lower case option with Stemming using Lovin's algorithm and setting minimum word frequency to 3.

An important note is that the weak performance of IDF Inverse Document Frequency option which could be due to the nature of the dataset where the documents are only tweets with very small amount of words 120 maximum.

Another Important note is that more options ticked doesn't result in better performance as seen in the table.

Chapter 5

Revisiting our classifiers

I know this was not required from this TP, but i wanted to retest all the classifiers again with the options that helped us the most in chapter 4. This ofcourse will be more buyest towards the classifier that we used in our tests which is Multinomial Naive Bayes.

5.1 C4.5 Default Settings

5.1.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         304               30.4    %
Incorrectly Classified Instances       696               69.6    %
Kappa statistic                          0.0928
Mean absolute error                      0.1129
Root mean squared error                  0.2806
Relative absolute error                 92.3007 %
Root relative squared error            113.5792 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.016    0.000      0.000    0.000      -0.019   0.479     0.019     empty
                 0.354    0.221    0.343      0.354    0.348      0.131    0.585     0.313     sadness
                 0.354    0.258    0.357      0.354    0.355      0.096    0.558     0.351     worry
                 0.000    0.006    0.000      0.000    0.000      -0.010   0.450     0.017     fun
                 0.426    0.337    0.279      0.426    0.337      0.078    0.586     0.271     neutral
                 0.130    0.025    0.226      0.130    0.165      0.136    0.618     0.117     hate
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.522     0.014     enthusiasm
                 0.241    0.010    0.412      0.241    0.304      0.300    0.642     0.136     love
                 0.021    0.020    0.050      0.021    0.029      0.001    0.546     0.054     surprise
                 0.000    0.009    0.000      0.000    0.000      -0.016   0.422     0.029     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.452     0.006     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.536     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.576     0.005     anger
Weighted Avg.    0.304    0.212    ?          0.304    ?          ?        0.568     0.257

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   3   7   0   9   0   0   0   1   1   0   0   0 |   a = empty
   5  87  62   2  73   8   1   7   0   1   0   0   0 |   b = sadness
   6  66 102   1  88  10   1   2   9   3   0   0   0 |   c = worry
   0   1   8   0   7   0   0   0   0   0   0   0   0 |   d = fun
   2  59  60   1 100   3   1   0   7   2   0   0   0 |   e = neutral
   0   8  18   1  18   7   0   0   1   1   0   0   0 |   f = hate
   0   1   3   0  10   0   0   0   0   0   0   0   0 |   g = enthusiasm
   1   9   4   0   7   0   0   7   0   1   0   0   0 |   h = love
   1   8   9   0  27   2   0   0   1   0   0   0   0 |   i = surprise
   1  10   7   1   7   1   0   1   1   0   0   0   0 |   j = happiness
   0   0   2   0   4   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   3   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

5.1.2 Percentage Split (66%)

```
=== Summary ===
Correctly Classified Instances          91               26.7647 %
Incorrectly Classified Instances       249               73.2353 %
Kappa statistic                          0.0495
Mean absolute error                      0.119
Root mean squared error                  0.2856
Relative absolute error                 96.8668 %
Root relative squared error            114.8635 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.006    0.000      0.000    0.000      -0.014   0.657     0.052     empty
                 0.349    0.296    0.276      0.349    0.309      0.050    0.503     0.254     sadness
                 0.266    0.187    0.352      0.266    0.303      0.087    0.540     0.304     worry
                 0.000    0.027    0.000      0.000    0.000      -0.016   0.457     0.008     fun
                 0.474    0.393    0.264      0.474    0.339      0.069    0.537     0.247     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.721     0.180     hate
                 0.000    0.000    ?          0.000    ?          ?        0.313     0.012     enthusiasm
                 0.000    0.021    0.000      0.000    0.000      -0.030   0.333     0.039     love
                 0.000    0.018    0.000      0.000    0.000      -0.026   0.538     0.042     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.539     0.034     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.491     0.009     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.536     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.596     0.008     anger
Weighted Avg.    0.268    0.216    ?          0.268    ?          ?        0.533     0.220

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  3  1  0  7  0  0  0  0  0  0  0  0 |  a = empty
  0 29 15  3 31  0  0  5  0  0  0  0  0 |  b = sadness
  0 30 25  3 35  0  0  0  1  0  0  0  0 |  c = worry
  0  0  1  0  1  0  0  0  1  0  0  0  0 |  d = fun
  1 15 19  3 37  0  0  0  3  0  0  0  0 |  e = neutral
  0  6  4  0  9  0  0  2  0  0  0  0  0 |  f = hate
  0  2  2  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0 11  2  0  1  0  0  0  0  0  0  0  0 |  h = love
  0  4  0  0  8  0  0  0  0  0  0  0  0 |  i = surprise
  1  3  0  0  5  0  0  0  1  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.2 KNN K=1

5.2.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         227               22.7    %
Incorrectly Classified Instances       773               77.3    %
Kappa statistic                          0.0315
Mean absolute error                      0.1233
Root mean squared error                  0.3073
Relative absolute error                100.7798 %
Root relative squared error            124.3788 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.143    0.110    0.027      0.143    0.045      0.015    0.498     0.028     empty
                 0.293    0.208    0.314      0.293    0.303      0.087    0.559     0.291     sadness
                 0.191    0.163    0.322      0.191    0.240      0.034    0.506     0.303     worry
                 0.000    0.028    0.000      0.000    0.000      -0.022   0.493     0.016     fun
                 0.396    0.383    0.241      0.396    0.300      0.011    0.492     0.224     neutral
                 0.019    0.030    0.034      0.019    0.024      -0.015   0.476     0.055     hate
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.432     0.012     enthusiasm
                 0.000    0.002    0.000      0.000    0.000      -0.008   0.566     0.033     love
                 0.063    0.015    0.176      0.063    0.092      0.079    0.599     0.081     surprise
                 0.000    0.005    0.000      0.000    0.000      -0.012   0.465     0.030     happiness
                 0.000    0.006    0.000      0.000    0.000      -0.006   0.557     0.007     boredom
                 0.000    0.006    0.000      0.000    0.000      -0.008   0.339     0.008     relief
                 0.000    0.007    0.000      0.000    0.000      -0.005   0.237     0.003     anger
Weighted Avg.    0.227    0.194    0.237      0.227    0.221      0.036    0.515     0.221

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   3   3   2   2  10   0   0   0   0   1   0   0   0 |   a = empty
```

```
  20  72  38   6  96   9   1   1   1   0   1   1   0 |   b = sadness
  40  64  55   3 104  10   0   0   7   1   2   0   2 |   c = worry
   1   1   4   0   9   1   0   0   0   0   0   0   0 |   d = fun
  27  43  35  15  93   6   0   1   6   1   2   4   2 |   e = neutral
   1  12  16   2  18   1   0   0   0   1   1   0   2 |   f = hate
   3   3   2   0   6   0   0   0   0   0   0   0   0 |   g = enthusiasm
   1   8   3   0  14   1   1   0   0   1   0   0   0 |   h = love
   5  11   4   0  23   0   0   0   3   0   0   1   1 |   i = surprise
   5   9   6   0   7   1   1   0   0   0   0   0   0 |   j = happiness
   3   0   2   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   1   3   3   0   4   0   0   0   0   0   0   0   0 |   l = relief
   1   0   1   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

5.2.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         103               30.2941 %
Incorrectly Classified Instances       237               69.7059 %
Kappa statistic                          0.1046
Mean absolute error                      0.1173
Root mean squared error                  0.2972
Relative absolute error                 95.4962 %
Root relative squared error            119.5365 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.006    0.000      0.000    0.000      -0.014   0.583     0.047     empty
                 0.470    0.237    0.390      0.470    0.426      0.219    0.602     0.326     sadness
                 0.245    0.159    0.371      0.245    0.295      0.100    0.541     0.325     worry
                 0.000    0.006    0.000      0.000    0.000      -0.007   0.411     0.008     fun
                 0.487    0.385    0.273      0.487    0.350      0.087    0.528     0.254     neutral
                 0.095    0.069    0.083      0.095    0.089      0.025    0.541     0.071     hate
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.640     0.019     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.483     0.040     love
                 0.083    0.018    0.143      0.083    0.105      0.085    0.420     0.043     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.433     0.026     happiness
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.439     0.009     boredom
                 0.000    0.006    0.000      0.000    0.000      -0.009   0.559     0.019     relief
                 0.000    0.000    ?          0.000    ?          ?        0.578     0.009     anger
Weighted Avg.    0.303    0.196    ?          0.303    ?          ?        0.544     0.238

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  2  0  8  0  0  0  0  0  0  0  0 |  a = empty
  0 39 13  1 23  7  0  0  0  0  0  0  0 |  b = sadness
  2 22 23  0 35  6  0  0  4  0  1  1  0 |  c = worry
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  d = fun
  0 23  6  0 38  7  1  0  2  0  0  1  0 |  e = neutral
  0  5  6  1  7  2  0  0  0  0  0  0  0 |  f = hate
  0  0  2  0  2  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  4  2  0  7  1  0  0  0  0  0  0  0 |  h = love
  0  2  0  0  9  0  0  0  1  0  0  0  0 |  i = surprise
  0  2  4  0  3  1  0  0  0  0  0  0  0 |  j = happiness
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  2  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.3 KNN K=30

5.3.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         272               27.2    %
Incorrectly Classified Instances       728               72.8    %
Kappa statistic                          0.0465
Mean absolute error                      0.1235
Root mean squared error                  0.2517
Relative absolute error                100.9826 %
Root relative squared error            101.8885 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.478     0.023     empty
                 0.171    0.082    0.404      0.171    0.240      0.125    0.574     0.309     sadness
                 0.010    0.006    0.429      0.010    0.020      0.026    0.544     0.346     worry
                 0.000    0.000    ?          0.000    ?          ?        0.458     0.019     fun
                 0.966    0.865    0.255      0.966    0.404      0.136    0.593     0.280     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.404     0.044     hate
                 0.000    0.000    ?          0.000    ?          ?        0.456     0.020     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.532     0.038     love
                 0.000    0.000    ?          0.000    ?          ?        0.509     0.050     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.508     0.040     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.455     0.007     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.526     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.269     0.003     anger
Weighted Avg.    0.272    0.225    ?          0.272    ?          ?        0.547     0.250

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   0   0   0  21   0   0   0   0   0   0   0   0 |   a = empty
   0  42   2   0 202   0   0   0   0   0   0   0   0 |   b = sadness
   0  33   3   0 252   0   0   0   0   0   0   0   0 |   c = worry
   0   1   0   0  15   0   0   0   0   0   0   0   0 |   d = fun
   0   6   2   0 227   0   0   0   0   0   0   0   0 |   e = neutral
   0   6   0   0  48   0   0   0   0   0   0   0   0 |   f = hate
   0   1   0   0  13   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0   6   0   0  23   0   0   0   0   0   0   0   0 |   h = love
   0   5   0   0  43   0   0   0   0   0   0   0   0 |   i = surprise
   0   3   0   0  26   0   0   0   0   0   0   0   0 |   j = happiness
   0   0   0   0   6   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   0   0  10   0   0   0   0   0   0   0   0 |   l = relief
   0   0   0   0   3   0   0   0   0   0   0   0   0 |   m = anger
```

5.3.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances          85               25      %
Incorrectly Classified Instances       255               75      %
Kappa statistic                          0.021
Mean absolute error                      0.1221
Root mean squared error                  0.2512
Relative absolute error                 99.3903 %
Root relative squared error            101.0244 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.489     0.037     empty
                 0.373    0.268    0.310      0.373    0.339      0.099    0.571     0.326     sadness
                 0.011    0.000    1.000      0.011    0.021      0.088    0.569     0.355     worry
                 0.000    0.000    ?          0.000    ?          ?        0.653     0.018     fun
                 0.679    0.710    0.222      0.679    0.334      -0.028   0.560     0.281     neutral
                 0.000    0.000    ?          0.000    ?          ?        0.396     0.056     hate
                 0.000    0.000    ?          0.000    ?          ?        0.575     0.030     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.441     0.037     love
                 0.000    0.000    ?          0.000    ?          ?        0.523     0.038     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.403     0.026     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.647     0.038     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.624     0.041     relief
                 0.000    0.000    ?          0.000    ?          ?        0.760     0.018     anger
Weighted Avg.    0.250    0.228    ?          0.250    ?          ?        0.546     0.252

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  4  0  0  7  0  0  0  0  0  0  0  0 |  a = empty
  0 31  0  0 52  0  0  0  0  0  0  0  0 |  b = sadness
  0 21  1  0 72  0  0  0  0  0  0  0  0 |  c = worry
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  d = fun
  0 25  0  0 53  0  0  0  0  0  0  0  0 |  e = neutral
  0  7  0  0 14  0  0  0  0  0  0  0  0 |  f = hate
  0  2  0  0  2  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  5  0  0  9  0  0  0  0  0  0  0  0 |  h = love
  0  3  0  0  9  0  0  0  0  0  0  0  0 |  i = surprise
  0  1  0  0  9  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  1  0  0  4  0  0  0  0  0  0  0  0 |  l = relief
  0  0  0  0  2  0  0  0  0  0  0  0  0 |  m = anger
```

5.4 Logistic Regression Default Settings

5.4.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         359               35.9    %
Incorrectly Classified Instances       641               64.1    %
Kappa statistic                          0.1443
Mean absolute error                      0.1176
Root mean squared error                  0.2449
Relative absolute error                 96.1278 %
Root relative squared error             99.1476 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.007   0.703     0.039     empty
                 0.244    0.082    0.492      0.244    0.326      0.213    0.589     0.365     sadness
                 0.559    0.402    0.360      0.559    0.438      0.143    0.584     0.361     worry
                 0.000    0.004    0.000      0.000    0.000      -0.008   0.484     0.018     fun
                 0.540    0.337    0.330      0.540    0.410      0.177    0.654     0.316     neutral
                 0.111    0.011    0.375      0.111    0.171      0.181    0.583     0.128     hate
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.454     0.014     enthusiasm
                 0.172    0.006    0.455      0.172    0.250      0.267    0.739     0.186     love
                 0.000    0.004    0.000      0.000    0.000      -0.014   0.527     0.050     surprise
                 0.000    0.004    0.000      0.000    0.000      -0.011   0.472     0.029     happiness
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.598     0.012     boredom
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.395     0.009     relief
                 0.000    0.000    ?          0.000    ?          ?        0.220     0.002     anger
Weighted Avg.    0.359    0.216    ?          0.359    ?          ?        0.596     0.285

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   0   7   0  14   0   0   0   0   0   0   0   0 |   a = empty
   0  60  96   0  76   7   0   4   1   2   0   0   0 |   b = sadness
   1  37 161   0  82   2   1   1   1   1   1   0   0 |   c = worry
   0   1   7   0   6   0   1   0   0   1   0   0   0 |   d = fun
   1  11  89   3 127   0   0   1   2   0   0   1   0 |   e = neutral
   0   3  22   1  22   6   0   0   0   0   0   0   0 |   f = hate
   0   0   9   0   4   1   0   0   0   0   0   0   0 |   g = enthusiasm
   0   4  12   0   8   0   0   5   0   0   0   0   0 |   h = love
   0   1  20   0  27   0   0   0   0   0   0   0   0 |   i = surprise
   0   4  16   0   8   0   1   0   0   0   0   0   0 |   j = happiness
   0   0   4   0   2   0   0   0   0   0   0   0   0 |   k = boredom
   0   1   3   0   7   0   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

5.4.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         113               33.2353 %
Incorrectly Classified Instances       227               66.7647 %
Kappa statistic                          0.1368
Mean absolute error                      0.1139
Root mean squared error                  0.2543
Relative absolute error                 92.6836 %
Root relative squared error            102.2919 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.003    0.000      0.000    0.000      -0.010   0.741     0.104     empty
                 0.313    0.140    0.419      0.313    0.359      0.193    0.657     0.397     sadness
                 0.468    0.321    0.358      0.468    0.406      0.137    0.605     0.377     worry
                 0.000    0.009    0.000      0.000    0.000      -0.009   0.320     0.008     fun
                 0.513    0.290    0.345      0.513    0.412      0.198    0.652     0.325     neutral
                 0.000    0.013    0.000      0.000    0.000      -0.028   0.530     0.078     hate
                 0.000    0.006    0.000      0.000    0.000      -0.008   0.286     0.010     enthusiasm
                 0.214    0.015    0.375      0.214    0.273      0.261    0.721     0.207     love
                 0.000    0.024    0.000      0.000    0.000      -0.030   0.599     0.046     surprise
                 0.000    0.030    0.000      0.000    0.000      -0.030   0.530     0.042     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.028     boredom
                 0.000    0.009    0.000      0.000    0.000      -0.012   0.459     0.017     relief
                 0.000    0.000    ?          0.000    ?          ?        0.809     0.028     anger
Weighted Avg.    0.332    0.193    ?          0.332    ?          ?        0.624     0.296

=== Confusion Matrix ===
  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  3  0  6  0  0  0  0  1  0  0  0 |  a = empty
  1 26 31  1 16  1  2  1  2  1  0  1  0 |  b = sadness
  0 12 44  0 27  3  0  0  3  5  0  0  0 |  c = worry
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  d = fun
  0  7 23  0 40  0  0  2  3  1  0  2  0 |  e = neutral
  0  7  4  2  6  0  0  1  0  1  0  0  0 |  f = hate
  0  0  2  0  2  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  2  4  0  5  0  0  3  0  0  0  0  0 |  h = love
  0  1  1  0  9  0  0  1  0  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  0  0  0  0  0  1  0  0  0 |  m = anger
```

5.5 Naive Bayes Default Settings

5.5.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         283               28.3    %
Incorrectly Classified Instances       717               71.7    %
Kappa statistic                          0.1149
Mean absolute error                      0.1139
Root mean squared error                  0.2828
Relative absolute error                 93.0798 %
Root relative squared error            114.4787 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.048    0.026    0.038      0.048    0.043      0.020    0.650     0.031     empty
                 0.313    0.176    0.367      0.313    0.338      0.144    0.590     0.391     sadness
                 0.354    0.213    0.402      0.354    0.376      0.146    0.605     0.384     worry
                 0.063    0.023    0.042      0.063    0.050      0.032    0.489     0.019     fun
                 0.340    0.200    0.343      0.340    0.342      0.141    0.620     0.315     neutral
                 0.204    0.062    0.157      0.204    0.177      0.125    0.665     0.129     hate
                 0.000    0.021    0.000      0.000    0.000      -0.017   0.549     0.017     enthusiasm
                 0.103    0.045    0.064      0.103    0.079      0.046    0.652     0.056     love
                 0.083    0.050    0.077      0.083    0.080      0.032    0.592     0.066     surprise
                 0.138    0.042    0.089      0.138    0.108      0.077    0.660     0.066     happiness
                 0.000    0.002    0.000      0.000    0.000      -0.003   0.554     0.008     boredom
                 0.000    0.014    0.000      0.000    0.000      -0.013   0.322     0.009     relief
                 0.000    0.002    0.000      0.000    0.000      -0.002   0.739     0.011     anger
Weighted Avg.    0.283    0.162    0.305      0.283    0.292      0.123    0.606     0.296

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   1   0   8   1   9   0   0   0   1   0   0   1   0 |   a = empty
   5  77  52   2  55  18   1  12  10  11   0   3   0 |   b = sadness
   7  57 102   9  45  21   8  12   8  12   2   4   1 |   c = worry
   1   3   3   1   3   2   0   0   1   1   0   1   0 |   d = fun
   7  27  54   5  80   9  10   9  21   7   0   5   1 |   e = neutral
   2  13   9   2  11  11   0   2   1   3   0   0   0 |   f = hate
   0   2   5   1   3   1   0   1   0   1   0   0   0 |   g = enthusiasm
   1   8   3   0   5   2   0   3   4   3   0   0   0 |   h = love
   1   9  11   1  12   4   1   3   4   2   0   0   0 |   i = surprise
   0   9   2   2   4   2   0   4   2   4   0   0   0 |   j = happiness
   0   1   2   0   1   0   1   1   0   0   0   0   0 |   k = boredom
   1   4   2   0   3   0   0   0   0   1   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

5.5.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         101               29.7059 %
Incorrectly Classified Instances       239               70.2941 %
Kappa statistic                          0.1249
Mean absolute error                      0.1132
Root mean squared error                  0.2834
Relative absolute error                 92.1507 %
Root relative squared error            113.973  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.015    0.000      0.000    0.000      -0.022   0.752     0.077     empty
                 0.325    0.183    0.365      0.325    0.344      0.148    0.574     0.357     sadness
                 0.351    0.240    0.359      0.351    0.355      0.112    0.576     0.364     worry
                 0.000    0.018    0.000      0.000    0.000      -0.013   0.463     0.011     fun
                 0.436    0.210    0.382      0.436    0.407      0.216    0.670     0.330     neutral
                 0.143    0.038    0.200      0.143    0.167      0.123    0.649     0.146     hate
                 0.000    0.015    0.000      0.000    0.000      -0.013   0.539     0.016     enthusiasm
                 0.071    0.028    0.100      0.071    0.083      0.052    0.556     0.121     love
                 0.083    0.049    0.059      0.083    0.069      0.029    0.519     0.046     surprise
                 0.200    0.070    0.080      0.200    0.114      0.084    0.661     0.065     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.792     0.036     boredom
                 0.000    0.006    0.000      0.000    0.000      -0.009   0.372     0.015     relief
                 0.000    0.000    ?          0.000    ?          ?        0.717     0.017     anger
Weighted Avg.    0.297    0.167    ?          0.297    ?          ?        0.605     0.284

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  0  4  1  6  0  0  0  0  0  0  0  0 |  a = empty
  0 27 22  2 15  2  1  2  3  8  0  1  0 |  b = sadness
  2 16 33  1 16  6  2  3  5  9  0  1  0 |  c = worry
  0  0  0  0  1  0  0  0  2  0  0  0  0 |  d = fun
  2 12 18  0 34  2  1  0  5  4  0  0  0 |  e = neutral
  0  6  5  1  2  3  0  3  1  0  0  0  0 |  f = hate
  0  0  2  0  2  0  0  0  0  0  0  0  0 |  g = enthusiasm
  1  4  2  0  3  0  1  1  0  2  0  0  0 |  h = love
  0  1  1  1  6  1  0  1  1  0  0  0  0 |  i = surprise
  0  4  2  0  1  1  0  0  0  2  0  0  0 |  j = happiness
  0  1  1  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  3  1  0  1  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.6 Multinomial Naive Bayes Default Settings

5.6.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         374               37.4    %
Incorrectly Classified Instances       626               62.6    %
Kappa statistic                          0.1531
Mean absolute error                      0.1103
Root mean squared error                  0.2449
Relative absolute error                 90.1928 %
Root relative squared error             99.1119 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.029     empty
                 0.447    0.223    0.396      0.447    0.420      0.216    0.662     0.417     sadness
                 0.580    0.403    0.368      0.580    0.450      0.161    0.645     0.427     worry
                 0.000    0.000    ?          0.000    ?          ?        0.508     0.027     fun
                 0.413    0.214    0.372      0.413    0.391      0.192    0.656     0.331     neutral
                 0.000    0.005    0.000      0.000    0.000      -0.017   0.565     0.083     hate
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.018     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.523     0.055     love
                 0.000    0.002    0.000      0.000    0.000      -0.010   0.565     0.059     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.525     0.033     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.008     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.496     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.661     0.008     anger
Weighted Avg.    0.374    0.222    ?          0.374    ?          ?        0.629     0.315

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   1  15   0   5   0   0   0   0   0   0   0   0 |   a = empty
   0 110  90   0  41   4   0   0   1   0   0   0   0 |   b = sadness
   0  62 167   0  57   1   0   0   1   0   0   0   0 |   c = worry
   0   4   5   0   7   0   0   0   0   0   0   0   0 |   d = fun
   0  43  95   0  97   0   0   0   0   0   0   0   0 |   e = neutral
   0  17  26   0  11   0   0   0   0   0   0   0   0 |   f = hate
   0   1   8   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  19   4   0   6   0   0   0   0   0   0   0   0 |   h = love
   0  11  20   0  17   0   0   0   0   0   0   0   0 |   i = surprise
   0   7  14   0   8   0   0   0   0   0   0   0   0 |   j = happiness
   0   1   4   0   1   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   3   0   6   0   0   0   0   0   0   0   0 |   l = relief
   0   0   3   0   0   0   0   0   0   0   0   0   0 |   m = anger
```

5.6.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         118               34.7059 %
Incorrectly Classified Instances       222               65.2941 %
Kappa statistic                          0.1266
Mean absolute error                      0.1122
Root mean squared error                  0.2498
Relative absolute error                 91.3624 %
Root relative squared error            100.4543 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.666     0.054     empty
                 0.398    0.218    0.371      0.398    0.384      0.176    0.633     0.371     sadness
                 0.564    0.427    0.335      0.564    0.421      0.123    0.617     0.427     worry
                 0.000    0.000    ?          0.000    ?          ?        0.585     0.017     fun
                 0.410    0.214    0.364      0.410    0.386      0.189    0.628     0.317     neutral
                 0.000    0.006    0.000      0.000    0.000      -0.020   0.595     0.080     hate
                 0.000    0.000    ?          0.000    ?          ?        0.398     0.015     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.463     0.077     love
                 0.000    0.009    0.000      0.000    0.000      -0.018   0.635     0.066     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.547     0.042     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.639     0.043     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.418     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.593     0.012     anger
Weighted Avg.    0.347    0.221    ?          0.347    ?          ?        0.610     0.296

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  1  6  0  4  0  0  0  0  0  0  0  0 |  a = empty
  0 33 35  0 13  1  0  0  1  0  0  0  0 |  b = sadness
  0 19 53  0 21  1  0  0  0  0  0  0  0 |  c = worry
  0  0  2  0  0  0  0  0  1  0  0  0  0 |  d = fun
  0 12 33  0 32  0  0  0  1  0  0  0  0 |  e = neutral
  0  7 10  0  4  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  7  5  0  2  0  0  0  0  0  0  0  0 |  h = love
  0  3  2  0  7  0  0  0  0  0  0  0  0 |  i = surprise
  0  5  4  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  2  0  1  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.7 Random Forest Default Settings

5.7.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         317               31.7    %
Incorrectly Classified Instances       683               68.3    %
Kappa statistic                          0.0887
Mean absolute error                      0.1193
Root mean squared error                  0.2492
Relative absolute error                 97.549  %
Root relative squared error            100.8781 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.010    0.000      0.000    0.000      -0.015   0.591     0.030     empty
                 0.333    0.191    0.363      0.333    0.347      0.147    0.604     0.364     sadness
                 0.444    0.385    0.318      0.444    0.371      0.055    0.565     0.320     worry
                 0.000    0.009    0.000      0.000    0.000      -0.012   0.553     0.023     fun
                 0.447    0.295    0.317      0.447    0.371      0.136    0.587     0.286     neutral
                 0.000    0.008    0.000      0.000    0.000      -0.021   0.580     0.071     hate
                 0.000    0.000    ?          0.000    ?          ?        0.496     0.014     enthusiasm
                 0.000    0.004    0.000      0.000    0.000      -0.011   0.701     0.081     love
                 0.042    0.007    0.222      0.042    0.070      0.078    0.607     0.103     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.537     0.039     happiness
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.563     0.015     boredom
                 0.000    0.000    ?          0.000    ?          ?        0.517     0.012     relief
                 0.000    0.000    ?          0.000    ?          ?        0.446     0.003     anger
Weighted Avg.    0.317    0.229    ?          0.317    ?          ?        0.584     0.263

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   0   2   6   1  12   0   0   0   0   0   0   0   0 |   a = empty
   1  82  87   2  68   3   0   2   1   0   0   0   0 |   b = sadness
   4  73 128   1  78   1   0   0   3   0   0   0   0 |   c = worry
   0   1   7   0   7   1   0   0   0   0   0   0   0 |   d = fun
   4  31  85   4 105   3   0   1   2   0   0   0   0 |   e = neutral
   0  10  28   1  14   0   0   0   0   0   1   0   0 |   f = hate
   1   1   8   0   4   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0  10   9   0   9   0   0   0   1   0   0   0   0 |   h = love
   0   8  19   0  19   0   0   0   2   0   0   0   0 |   i = surprise
   0   6  15   0   7   0   0   1   0   0   0   0   0 |   j = happiness
   0   0   4   0   2   0   0   0   0   0   0   0   0 |   k = boredom
   0   2   4   0   5   0   0   0   0   0   0   0   0 |   l = relief
   0   0   2   0   1   0   0   0   0   0   0   0   0 |   m = anger
```

5.7.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         113               33.2353 %
Incorrectly Classified Instances       227               66.7647 %
Kappa statistic                          0.1186
Mean absolute error                      0.1183
Root mean squared error                  0.2478
Relative absolute error                 96.2806 %
Root relative squared error             99.6562 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.640     0.057     empty
                 0.410    0.233    0.362      0.410    0.384      0.169    0.609     0.359     sadness
                 0.426    0.264    0.381      0.426    0.402      0.156    0.612     0.358     worry
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.356     0.008     fun
                 0.487    0.355    0.290      0.487    0.364      0.114    0.624     0.338     neutral
                 0.000    0.016    0.000      0.000    0.000      -0.031   0.585     0.107     hate
                 0.000    0.000    ?          0.000    ?          ?        0.688     0.030     enthusiasm
                 0.000    0.000    ?          0.000    ?          ?        0.582     0.052     love
                 0.083    0.006    0.333      0.083    0.133      0.152    0.647     0.149     surprise
                 0.000    0.000    ?          0.000    ?          ?        0.580     0.037     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.563     0.014     boredom
                 0.000    0.003    0.000      0.000    0.000      -0.007   0.608     0.023     relief
                 0.000    0.000    ?          0.000    ?          ?        0.828     0.032     anger
Weighted Avg.    0.332    0.213    ?          0.332    ?          ?        0.612     0.282

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  2  0  0  9  0  0  0  0  0  0  0  0 |  a = empty
  0 34 23  0 25  1  0  0  0  0  0  0  0 |  b = sadness
  0 17 40  0 34  1  0  0  1  0  0  1  0 |  c = worry
  0  0  1  0  2  0  0  0  0  0  0  0  0 |  d = fun
  0 17 19  0 38  3  0  0  1  0  0  0  0 |  e = neutral
  0 11  8  1  1  0  0  0  0  0  0  0  0 |  f = hate
  0  0  4  0  0  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  6  2  0  6  0  0  0  0  0  0  0  0 |  h = love
  0  1  1  0  9  0  0  0  1  0  0  0  0 |  i = surprise
  0  4  5  0  1  0  0  0  0  0  0  0  0 |  j = happiness
  0  0  0  0  3  0  0  0  0  0  0  0  0 |  k = boredom
  0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
  0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.8 Support Vector Machine Default Settings

5.8.1 Cross Validation (10 Folds)

```
=== Summary ===

Correctly Classified Instances         339               33.9    %
Incorrectly Classified Instances       661               66.1    %
Kappa statistic                          0.1393
Mean absolute error                      0.1344
```

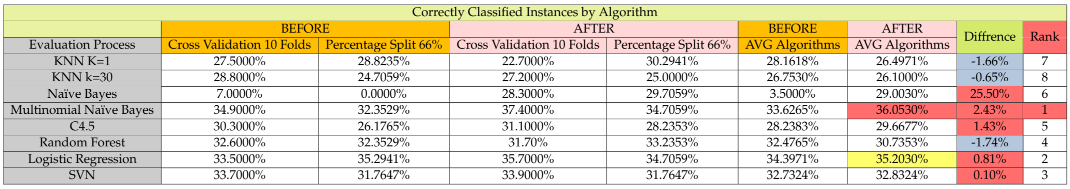
```
Root mean squared error                  0.2562
Relative absolute error                109.8751 %
Root relative squared error            103.7    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.048    0.018    0.053      0.048    0.050      0.031    0.699     0.036     empty
                 0.476    0.265    0.369      0.476    0.416      0.195    0.611     0.322     sadness
                 0.389    0.250    0.386      0.389    0.388      0.139    0.615     0.360     worry
                 0.000    0.007    0.000      0.000    0.000      -0.011   0.538     0.020     fun
                 0.438    0.237    0.363      0.438    0.397      0.190    0.639     0.320     neutral
                 0.037    0.021    0.091      0.037    0.053      0.024    0.519     0.060     hate
                 0.000    0.006    0.000      0.000    0.000      -0.009   0.432     0.013     enthusiasm
                 0.069    0.011    0.154      0.069    0.095      0.085    0.634     0.074     love
                 0.021    0.025    0.040      0.021    0.027      -0.006   0.582     0.059     surprise
                 0.034    0.014    0.067      0.034    0.045      0.028    0.508     0.041     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.467     0.006     boredom
                 0.000    0.002    0.000      0.000    0.000      -0.005   0.514     0.011     relief
                 0.000    0.000    ?          0.000    ?          ?        0.198     0.003     anger
Weighted Avg.    0.339    0.197    ?          0.339    ?          ?        0.605     0.269

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
   1   6   6   0   8   0   0   0   0   0   0   0   0 |   a = empty
   6 117  60   0  43   4   0   5   7   4   0   0   0 |   b = sadness
   6  85 112   3  67   5   1   1   4   3   0   1   0 |   c = worry
   0   1   5   0   6   2   0   0   1   1   0   0   0 |   d = fun
   4  54  55   1 103   2   5   3   4   3   0   1   0 |   e = neutral
   0  16  20   1  13   2   0   0   1   1   0   0   0 |   f = hate
   0   5   4   0   5   0   0   0   0   0   0   0   0 |   g = enthusiasm
   0   7   5   1   7   1   0   2   4   2   0   0   0 |   h = love
   1  11  11   1  19   3   0   1   1   0   0   0   0 |   i = surprise
   0  11   6   0   6   1   0   1   3   1   0   0   0 |   j = happiness
   0   1   2   0   2   1   0   0   0   0   0   0   0 |   k = boredom
   1   3   3   0   3   1   0   0   0   0   0   0   0 |   l = relief
   0   0   1   0   2   0   0   0   0   0   0   0   0 |   m = anger
```

5.8.2 Percentage Split (66%)

```
=== Summary ===

Correctly Classified Instances         107               31.4706 %
Incorrectly Classified Instances       233               68.5294 %
Kappa statistic                          0.1164
Mean absolute error                      0.1347
Root mean squared error                  0.2568
Relative absolute error                109.6506 %
Root relative squared error            103.2874 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.018    0.000      0.000    0.000      -0.025   0.729     0.101     empty
                 0.410    0.241    0.354      0.410    0.380      0.161    0.595     0.313     sadness
                 0.447    0.289    0.372      0.447    0.406      0.150    0.571     0.317     worry
                 0.000    0.012    0.000      0.000    0.000      -0.010   0.460     0.009     fun
                 0.372    0.225    0.330      0.372    0.349      0.141    0.638     0.302     neutral
                 0.048    0.013    0.200      0.048    0.077      0.070    0.526     0.089     hate
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.682     0.057     enthusiasm
                 0.000    0.009    0.000      0.000    0.000      -0.020   0.469     0.039     love
                 0.083    0.046    0.063      0.083    0.071      0.033    0.620     0.051     surprise
                 0.000    0.015    0.000      0.000    0.000      -0.021   0.500     0.036     happiness
                 0.000    0.000    ?          0.000    ?          ?        0.457     0.009     boredom
                 0.000    0.009    0.000      0.000    0.000      -0.012   0.488     0.016     relief
                 0.000    0.000    ?          0.000    ?          ?        0.527     0.006     anger
Weighted Avg.    0.315    0.194    ?          0.315    ?          ?        0.588     0.248

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
  0  4  2  0  5  0  0  0  0  0  0  0  0 |  a = empty
  0 34 27  1 13  2  0  1  3  1  0  1  0 |  b = sadness
  3 19 42  2 19  2  0  0  4  2  0  1  0 |  c = worry
  0  0  1  0  1  0  0  0  1  0  0  0  0 |  d = fun
  3 15 21  0 29  0  1  1  6  1  0  1  0 |  e = neutral
  0 10  6  1  3  1  0  0  0  0  0  0  0 |  f = hate
  0  1  2  0  1  0  0  0  0  0  0  0  0 |  g = enthusiasm
  0  4  4  0  4  0  0  0  1  1  0  0  0 |  h = love
  0  2  3  0  5  0  0  1  1  0  0  0  0 |  i = surprise
  0  4  3  0  3  0  0  0  0  0  0  0  0 |  j = happiness
```

```
 0  1  0  0  2  0  0  0  0  0  0  0  0 |  k = boredom
 0  2  1  0  2  0  0  0  0  0  0  0  0 |  l = relief
 0  0  1  0  1  0  0  0  0  0  0  0  0 |  m = anger
```

5.9 Conclusion



Correctly Classified Instances by Algorithm								
	BEFORE		AFTER		BEFORE	AFTER	Diffrence	Rank
Evaluation Process	Cross Validation 10 Folds	Percentage Split 66%	Cross Validation 10 Folds	Percentage Split 66%	AVG Algorithms	AVG Algorithms		
KNN K=1	27.5000%	28.8235%	22.7000%	30.2941%	28.1618%	26.4971%	-1.66%	7
KNN k=30	28.8000%	24.7059%	27.2000%	25.0000%	26.7530%	26.1000%	-0.65%	8
Naïve Bayes	7.0000%	0.0000%	28.3000%	29.7059%	3.5000%	29.0030%	25.50%	6
Multinomial Naïve Bayes	34.9000%	32.3529%	37.4000%	34.7059%	33.6265%	36.0530%	2.43%	1
C4.5	30.3000%	26.1765%	31.1000%	28.2353%	28.2383%	29.6677%	1.43%	5
Random Forest	32.6000%	32.3529%	31.70%	33.2353%	32.4765%	30.7353%	-1.74%	4
Logistic Regression	33.5000%	35.2941%	35.7000%	34.7059%	34.3971%	35.2030%	0.81%	2
SVN	33.7000%	31.7647%	33.9000%	31.7647%	32.7324%	32.8324%	0.10%	3

Now it must be noted that these options on the tests were selected based on Multinomial Naive Bayes performances in chapter 4. so they are more tailored for it than other algorithms. so it is normal that it was first. however the importance here is to see what will happen to the other algorithms.

The surprise was that some had a decrease in their performance which was not expected like KNN with both K=1 and K=30. Random Forest also had a decrease of 1.74%.

The other surprise was Naive Bayes which benefited the most with a huge 25.5% jump in performance.

However looking at the big picture we see that most algorithms didn't change much and even Multinomial Naive Bayes had its best with only 2.43% improvement only. this means that the best classifiers 1. Multinomial Naive Bayes, 2.Logistic Regression and 3.Support Vector Machine, 4.Random Forest, 5. C4.5 are very effective in Text Classification and don't need much help which is great as sometimes removing stop words or stemming can affect the general meaning of a sentence and in some cases it is better to keep them rather than removing or modifying them. and a general rule the best dataset is the one that is as pure as possible in order to not affect the results in any way.

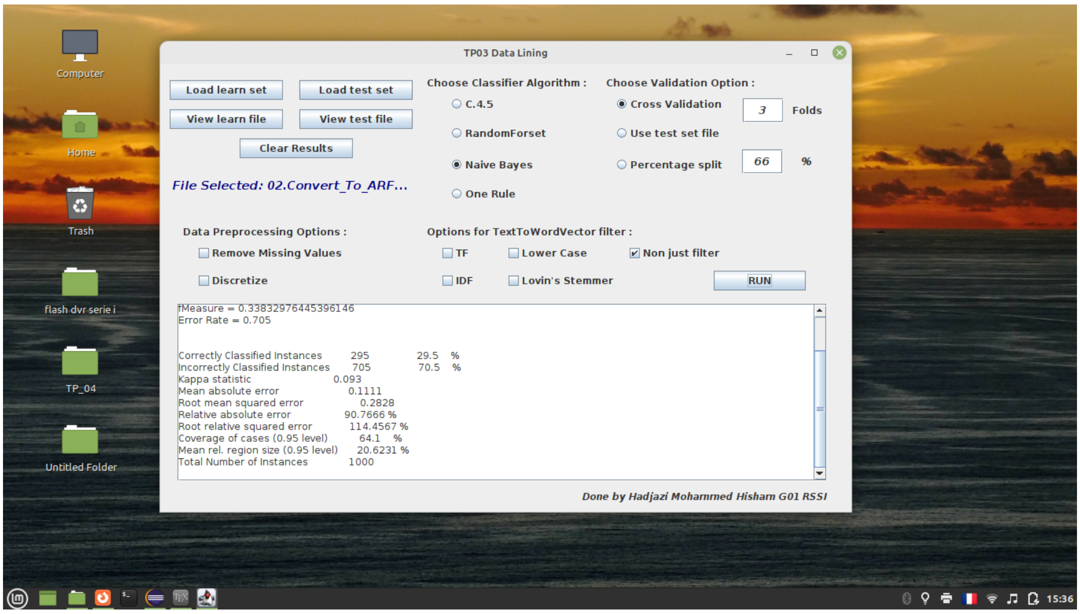
Chapter 6

Coding part

6.1 Execution with StringToWordVector and no Options

I didn't have enough time to do all the options I wanted to do in the **StringToWordVector()** filter, so I was only able to implement 2 options along with the filter only with no options. the options I have implemented are Lowercase option and Stemmer option with lovin's algorithm selected. still that is more than what is required from us which is one filter only with one classifier only.

```
        StringToWordVector filter = new StringToWordVector();
        filter.setInputFormat(datasetInstances);
        filter.setLowerCaseTokens(false);
        filter.setTFTransform(false);
        filter.setIDFTransform(false);
        datasetInstances = Filter.useFilter(datasetInstances,
            filter);
        datasetInstances.setClassIndex(0);
```



File Selected: 02.Convert_To_ARF...

Choose Classifier Algorithm :
○ C.4.5
○ RandomForset
● Naive Bayes
○ One Rule

Choose Validation Option :
● Cross Validation 3 Folds
○ Use test set file
○ Percentage split 66 %

Data Preprocessing Options :
□ Remove Missing Values
□ Discretize

Options for TextToWordVector filter :
□ TF
□ IDF
□ Lower Case
□ Lovin's Stemmer
☑ Non just filter

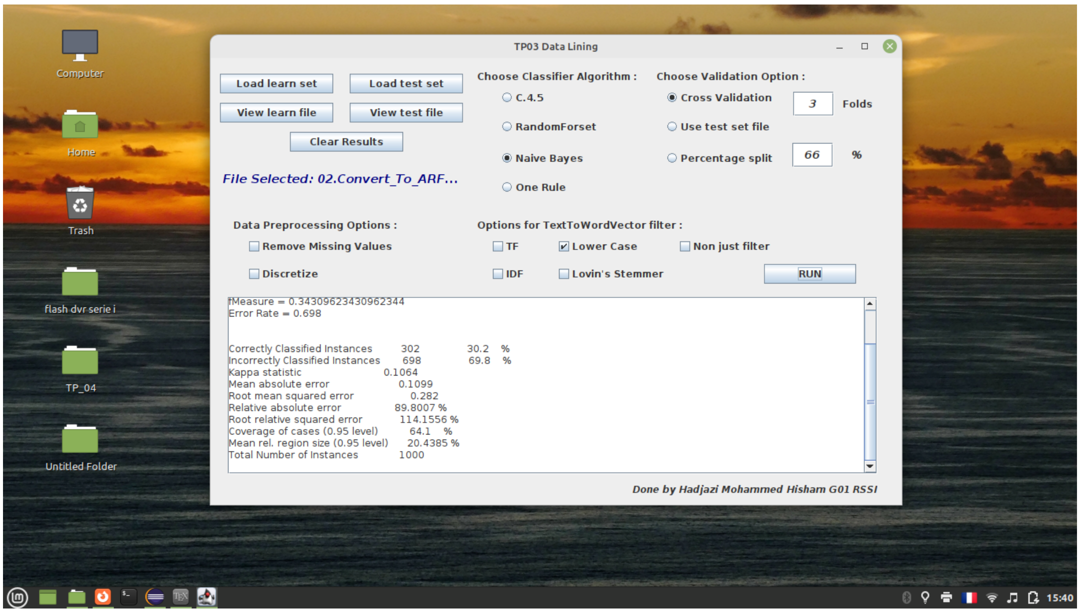
```
fMeasure = 0.33832976445396146
Error Rate = 0.705


Correctly Classified Instances        295               29.5    %
Incorrectly Classified Instances      705               70.5    %
Kappa statistic                     0.093
Mean absolute error                    0.1111
Root mean squared error                  0.2828
Relative absolute error             90.7666 %
Root relative squared error          114.4567 %
Coverage of cases (0.95 level)         64.1    %
Mean rel. region size (0.95 level)    20.6231 %
Total Number of Instances            1000
```

Done by Hadjazi Mohammed Hisham G01 RSSI

FIGURE 6.1: No Filters

6.2 Using Filters

```
            StringToWordVector filter = new StringToWordVector();
            filter.setInputFormat(datasetInstances);
            filter.setLowerCaseTokens(true);
            filter.setTFTransform(true);
            filter.setIDFTransform(true);
            LovinsStemmer stemmer = new LovinsStemmer();
            filter.setStemmer(stemmer);
            datasetInstances = Filter.useFilter(datasetInstances,
               filter);
            datasetInstances.setClassIndex(0);
```



File Selected: 02.Convert_To_ARF...

Choose Classifier Algorithm :
○ C.4.5
○ RandomForset
● Naive Bayes
○ One Rule

Choose Validation Option :
● Cross Validation 3 Folds
○ Use test set file
○ Percentage split 66 %

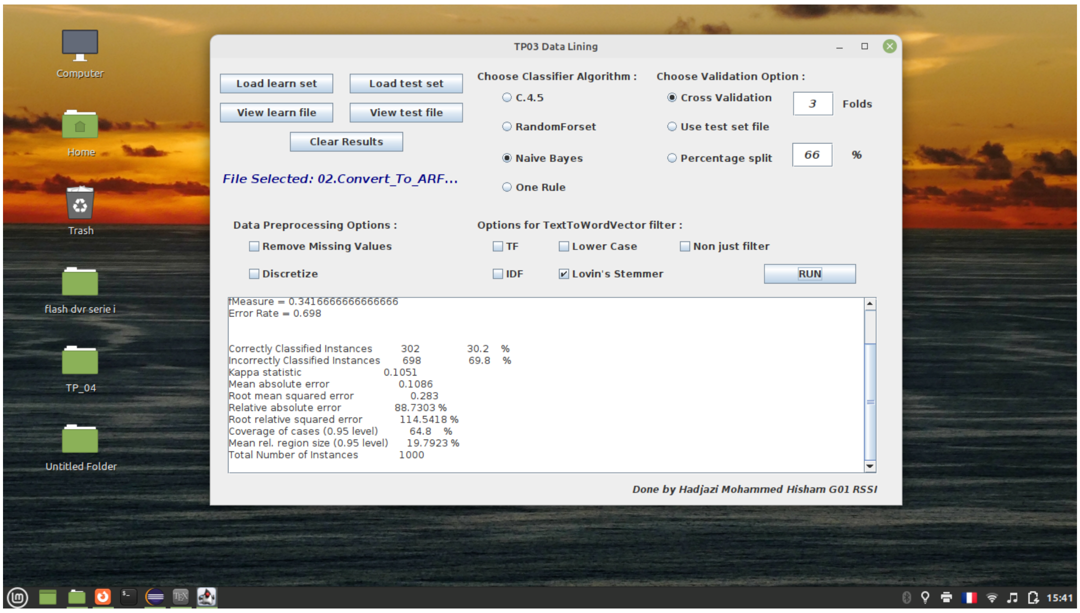
Data Preprocessing Options :
□ Remove Missing Values
□ Discretize

Options for TextToWordVector filter :
□ TF ☑ Lower Case □ Non just filter
□ IDF □ Lovin's Stemmer

```
fMeasure = 0.34309623430962344
Error Rate = 0.698


Correctly Classified Instances       302            30.2   %
Incorrectly Classified Instances     698            69.8   %
Kappa statistic                    0.1064
Mean absolute error                  0.1099
Root mean squared error                 0.282
Relative absolute error             89.8007 %
Root relative squared error          114.1556 %
Coverage of cases (0.95 level)        64.1    %
Mean rel. region size (0.95 level)    20.4385 %
Total Number of Instances            1000
```

Done by Hadjazi Mohammed Hisham G01 RSSI

FIGURE 6.2: Lower Case



File Selected: 02.Convert_To_ARF...

Choose Classifier Algorithm :
○ C.4.5
○ RandomForset
● Naive Bayes
○ One Rule

Choose Validation Option :
● Cross Validation 3 Folds
○ Use test set file
○ Percentage split 66 %

Data Preprocessing Options :
□ Remove Missing Values
□ Discretize

Options for TextToWordVector filter :
□ TF
□ Lower Case
□ Non just filter
□ IDF
☑ Lovin's Stemmer

```
fMeasure = 0.3416666666666666
Error Rate = 0.698


Correctly Classified Instances       302             30.2   %
Incorrectly Classified Instances     698             69.8   %
Kappa statistic                    0.1051
Mean absolute error                  0.1086
Root mean squared error                0.283
Relative absolute error             88.7303 %
Root relative squared error          114.5418 %
Coverage of cases (0.95 level)         64.8    %
Mean rel. region size (0.95 level)    19.7923 %
Total Number of Instances            1000
```

Done by Hadjazi Mohammed Hisham G01 RSSI

FIGURE 6.3: Lovins Stemmer

Bibliography

[1] Rohith Gandhi. *Support Vector Machine — Introduction to Machine Learning Algorithms*. en. July 2018. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (visited on 04/16/2022).

[2] Yassine Hamdaoui. *TF(Term Frequency)-IDF(Inverse Document Frequency) from scratch in python* . en. Mar. 2021. URL: https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558 (visited on 04/17/2022).

[3] Md Zahidul Islam et al. "A Semantics Aware Random Forest for Text Classification". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 1061–1070. ISBN: 9781450369763. DOI: 10.1145/3357384.3357891. URL: https://doi.org/10.1145/3357384.3357891 (visited on 04/16/2022).

[4] Idil Ismiguzel. *Applying Text Classification using Logistic Regression: A comparison between BoW and Tf-Idf*. en. Sept. 2021. URL: https://medium.com/analytics-vidhya/applying-text-classification-using-logistic-regression-a-comparison-between-bow-and-tf-idf-1f1ed1b83640 (visited on 04/16/2022).

[5] Kyung-Soon Lee and Kyo Kageura. "Virtual relevant documents in text categorization with support vector machines". en. In: *Information Processing & Management* 43.4 (July 2007), pp. 902–913. ISSN: 03064573. DOI: 10.1016/j.ipm.2006.08.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306457306001403 (visited on 04/16/2022).

[6] Susan Li. *Multi-Class Text Classification Model Comparison and Selection*. en. Dec. 2018. URL: https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568 (visited on 04/16/2022).

[7] *Machine Learning NLP Text Classification Algorithms and Models*. en. URL: https://www.projectpro.io/article/machine-learning-nlp-text-classification-algorithms-and-models/523 (visited on 04/16/2022).

[8] *Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022*. en-US. Jan. 2021. URL: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/ (visited on 04/16/2022).

[9] Rucha Narkhede and Rahul Gaikwad. "Spam Detection in Online Social Network with K-Means Clustering and SVM Machine Learning Approach". In: *International Journal of Innovations in Engineering and Science* 6.10 (Aug. 2021), p. 41. ISSN: 2456-3463. DOI: 10.46335/IJIES.2021.6.10.8. URL: http://ijies.net/finial-docs/finial-pdf/1308216108.pdf (visited on 04/16/2022).

[10] Syed Sadat Nazrul. *Multinomial Naive Bayes Classifier for Text Analysis (Python)*. en. June 2018. URL: https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67 (visited on 04/16/2022).

[11] Phu Vo Ngoc et al. "A C4.5 algorithm for english emotional classification". en. In: *Evolving Systems* 10.3 (Sept. 2019), pp. 425–451. ISSN: 1868-6486. DOI: 10.1007/s12530-017-9180-1. URL: https://doi.org/10.1007/s12530-017-9180-1 (visited on 04/16/2022).

[12] Christophe Pere. *Model Selection in Text Classification*. en. Nov. 2020. URL: https://towardsdatascience.com/model-selection-in-text-classification-ac13eedf6146 (visited on 04/16/2022).

[13] *sklearn.naive_bayes.MultinomialNB*. en. URL: https://scikit-learn/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (visited on 04/16/2022).

[14] *sklearn.naive_bayes.MultinomialNB*. en. URL: https://scikit-learn/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (visited on 04/16/2022).

[15] stefanoscerra. *Movie reviews classification using WEKA - a data mining experiment*. en-US. Sept. 2014. URL: https://www.stefanoscerra.it/movie-reviews-classification-weka-data-mining/ (visited on 04/16/2022).

[16] *Support Vector Machines (SVM) Algorithm Explained*. en. June 2017. URL: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/ (visited on 04/16/2022).

[17] *Text classification using K Nearest Neighbors (KNN)*. en. Dec. 2019. URL: https://iq.opengenus.org/text-classification-using-k-nearest-neighbors/ (visited on 04/16/2022).

[18] *The Lovins stemming algorithm*. URL: http://snowball.tartarus.org/algorithms/lovins/stemmer.html (visited on 04/17/2022).

[19] *What is Logistic regression? | IBM*. en-us. URL: https://www.ibm.com/topics/logistic-regression (visited on 04/16/2022).