R script for question 3

maelrazavet — *Mar 10, 2014, 6:47 PM*

```
setwd('~/Desktop//Courses/Warwick//Data Mining/Exercise Week 8/')
library(knitr)

require(e1071)

Loading required package: e1071
Loading required package: class

require(randomForest)

Loading required package: randomForest
randomForest 4.6-7
Type rfNews() to see new features/changes/bug fixes.

data = read.csv('AI2013_papers.csv')

d <- data
d <- d[,-1]
row.names(d) <- NULL
```

Let's normalize the data so that all features have a value between 0 and 1.

```
myNormalise=function(data, col){
  tmp = sqrt(sum(data[,col]^2))
  for(i in 1:nrow(data)){
    data[i,col] = data[i,col]/tmp
  }
  return(data)
}

for(attribute in 1:length(names(d))){
  if(is.numeric(d[,attribute]))
  {
    d <- myNormalise(d, attribute)
  }
}
```

Then, I need to implement the functions to use the three algorithms that we want to compare (Naïve Bayes, SVM and Random Forest). Note that in the following functions, we need to specify the current fold number. Indeed, we will implement the 10-fold cross-validation over the three algorithms. Therefore, each classifier is trained with the 9 other folds and tested on the current fold.

```
classes <- c('Case Study', 'Correspondence', 'Essay', 'Opinion', '
   Perspective', 'Research', 'Review', 'Viewpoint')
```

```
#function that implements the three algorithms to be compared
func_NB=function(fold, data){
  fitNaive = naiveBayes(type ~., data=data[data$fold != fold,-13])
  predictionsNaive = predict(fitNaive, data[data$fold == fold, -c(12,13)])
  return(table(predictionsNaive, data[data$fold==fold,]$type))
}
func_SVM=function(fold, data){
  fitSVM <- svm(type ~., data = data[data$fold != fold,-13])
  predictionsSVM = predict(fitSVM, data[data$fold == fold, -c(12,13)])
  return(table(predictionsSVM, data[data$fold==fold,]$type))
}
func_RF=function(fold, data){
  fitRandomForest <- randomForest(type ~., data = data[data$fold != fold
     ,-13], importance=TRUE)
  predictionsRF = predict(fitRandomForest, data[data$fold == fold, -c
     (12,13)])
  return(table(predictionsRF, data[data$fold==fold,]$type))
}
```

The next step is to implement a function that will compute the following measures for each fold and each classifier: Precision, Recall, F-score, Accuracy, the macro average precision and recall, and the micro average precision and recall. The accuracy will be then very useful to compare the three algorithms by applying the paired t-test.

```
#get Precision, Recall, F-measure, Accuracy, Rmacro, Rmicro, Pmacro and
   Pmicro for a fold for a classifier
getPRF = function(table){
  tab.precision <- c()
  tab.recall <- c()
  fmeasure <- c()
  acc <- c()
  Rmacro <- 0
  Rmicro <- 0
  Pmicro <- 0
  Pmacro <- 0
  TPc <- 0
  for(i in 1:ncol(table)){
    TP <- table[i,i]
    FN <- sum(table[-i,i])
    FP <- sum(table[i,-i])
    TN <- sum(table[-i,-i])
    precision <- TP/(TP+FP)
    recall <- TP/(TP+FN)
    f <- 2*precision*recall/(precision+recall)
    accuracy <- (TP + TN)/(TP + TN + FP + FN)

    tab.precision <- append(tab.precision, precision, after=length(tab.
       precision))
    tab.recall <- append(tab.recall, recall, after=length(tab.recall))
    fmeasure <- append(fmeasure, f, after=length(fmeasure))
    acc <- append(acc, accuracy, after=length(acc))

    Rmacro <- Rmacro + recall
```

```
    Pmacro <- Pmacro + precision
    TPc <- TPc + TP
    Rmicro <- TP + FN
    Pmicro <- TP + FP
  }
  Rmacro <- Rmacro/ncol(table)
  Pmacro <- Pmacro/ncol(table)
  Rmicro <- TPc / Rmicro
  Pmicro <- TPc / Pmicro
  return(data.frame(tab.precision, tab.recall, fmeasure, acc, Rmacro,
     Pmacro, Rmicro, Pmicro, row.names=classes))
}
```

The next step is to implement the function that will use the 10-fold cross-validation for a classifier.

```
#function that implements the cross-validation over a specific algorithm
doCV = function(data, kfold, algo){
  l=list()
  for (i in 1:kfold) {
    df = as.data.frame(getPRF(algo(i, data)))
    df[is.na(df)] <- 0
    l[[i]] = df
  }
  return (l)
}
d2<-d[sample(nrow(d)),]
d2$fold = cut(1:nrow(d2), breaks=10, labels=F)
row.names(d2) <- NULL
```

Let's have a look at the output of these functions, which are given below. Note that for each classifier, I display the different measures for each class and for each fold.

Results of the Naive Bayes classifier per fold

```
lNaive = doCV(d2, 10, func_NB)
lNaive
```

```
[[1]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.3636     0.5714   0.4444 0.9219  0.517      0
Correspondence        0.7059     1.0000   0.8276 0.9609  0.517      0
Essay                 0.3929     0.5500   0.4583 0.7969  0.517      0
Opinion               0.2564     0.9091   0.4000 0.7656  0.517      0
Perspective           0.0000     0.0000   0.0000 0.8281  0.517      0
Research              0.8421     0.7619   0.8000 0.9375  0.517      0
Review                0.9167     0.3438   0.5000 0.8281  0.517      0
Viewpoint             0.0000     0.0000   0.0000 0.9609  0.517      0
               Rmicro Pmicro
Case Study      21.33     32
Correspondence  21.33     32
Essay           21.33     32
Opinion         21.33     32
Perspective     21.33     32
```

```
Research         21.33     32
Review           21.33     32
Viewpoint        21.33     32

[[2]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.3478     0.5714   0.4324 0.8385 0.4755 0.5093
Correspondence        0.6364     1.0000   0.7778 0.9692 0.4755 0.5093
Essay                 0.5185     0.5600   0.5385 0.8154 0.4755 0.5093
Opinion               0.1212     0.5000   0.1951 0.7462 0.4755 0.5093
Perspective           0.2857     0.1250   0.1739 0.8538 0.4755 0.5093
Research              0.9000     0.5294   0.6667 0.9308 0.4755 0.5093
Review                0.7647     0.3514   0.4815 0.7846 0.4755 0.5093
Viewpoint             0.5000     0.1667   0.2500 0.9538 0.4755 0.5093
               Rmicro Pmicro
Case Study      9.667     29
Correspondence  9.667     29
Essay           9.667     29
Opinion         9.667     29
Perspective     9.667     29
Research        9.667     29
Review          9.667     29
Viewpoint       9.667     29

[[3]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.2727    0.20000  0.23077 0.8450 0.4715 0.4309
Correspondence        0.6667    1.00000  0.80000 0.9535 0.4715 0.4309
Essay                 0.3684    0.46667  0.41176 0.8450 0.4715 0.4309
Opinion               0.2927    0.92308  0.44444 0.7674 0.4715 0.4309
Perspective           0.2000    0.05556  0.08696 0.8372 0.4715 0.4309
Research              0.7647    0.59091  0.66667 0.8992 0.4715 0.4309
Review                0.8824    0.53571  0.66667 0.8837 0.4715 0.4309
Viewpoint             0.0000    0.00000  0.00000 0.9457 0.4715 0.4309
               Rmicro Pmicro
Case Study       10.5     63
Correspondence   10.5     63
Essay            10.5     63
Opinion          10.5     63
Perspective      10.5     63
Research         10.5     63
Review           10.5     63
Viewpoint        10.5     63

[[4]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.2667    0.28571   0.2759 0.8372 0.4142 0.4481
Correspondence        0.3077    1.00000   0.4706 0.9302 0.4142 0.4481
Essay                 0.4242    0.56000   0.4828 0.7674 0.4142 0.4481
Opinion               0.1212    0.57143   0.2000 0.7519 0.4142 0.4481
Perspective           0.4000    0.09524   0.1538 0.8295 0.4142 0.4481
Research              0.5652    0.56522   0.5652 0.8450 0.4142 0.4481
Review                0.5000    0.11111   0.1818 0.7907 0.4142 0.4481
Viewpoint             1.0000    0.12500   0.2222 0.9457 0.4142 0.4481
```

```
                Rmicro Pmicro
Case Study       5.625     45
Correspondence   5.625     45
Essay            5.625     45
Opinion          5.625     45
Perspective      5.625     45
Research         5.625     45
Review           5.625     45
Viewpoint        5.625     45

[[5]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.27273     0.3000   0.2857 0.8837 0.4702 0.3926
Correspondence        0.73333     1.0000   0.8462 0.9690 0.4702 0.3926
Essay                 0.35714     0.4545   0.4000 0.7674 0.4702 0.3926
Opinion               0.06667     1.0000   0.1250 0.7829 0.4702 0.3926
Perspective           0.26667     0.2222   0.2424 0.8062 0.4702 0.3926
Research              0.77778     0.6087   0.6829 0.8992 0.4702 0.3926
Review                0.66667     0.1765   0.2791 0.7597 0.4702 0.3926
Viewpoint             0.00000     0.0000   0.0000 0.9070 0.4702 0.3926
                Rmicro Pmicro
Case Study       5.556  16.67
Correspondence   5.556  16.67
Essay            5.556  16.67
Opinion          5.556  16.67
Perspective      5.556  16.67
Research         5.556  16.67
Review           5.556  16.67
Viewpoint        5.556  16.67

[[6]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.42857     0.4000   0.4138 0.8682 0.3642 0.3357
Correspondence        0.48148     0.8667   0.6190 0.8760 0.3642 0.3357
Essay                 0.26471     0.6000   0.3673 0.7597 0.3642 0.3357
Opinion               0.08696     0.2857   0.1333 0.7984 0.3642 0.3357
Perspective           0.20000     0.0500   0.0800 0.8217 0.3642 0.3357
Research              0.45455     0.4545   0.4545 0.9070 0.3642 0.3357
Review                0.76923     0.2564   0.3846 0.7519 0.3642 0.3357
Viewpoint             0.00000     0.0000   0.0000 0.9302 0.3642 0.3357
                Rmicro Pmicro
Case Study       6.571     23
Correspondence   6.571     23
Essay            6.571     23
Opinion          6.571     23
Perspective      6.571     23
Research         6.571     23
Review           6.571     23
Viewpoint        6.571     23

[[7]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study             0.4000     0.2222   0.2857 0.9225 0.4137      0
Correspondence         0.5333     0.8000   0.6400 0.9302 0.4137      0
```

```
Essay                0.2973      0.5789    0.3929 0.7364 0.4137      0
Opinion              0.1111      0.5000    0.1818 0.7209 0.4137      0
Perspective          0.0000      0.0000    0.0000 0.7829 0.4137      0
Research             0.8947      0.7083    0.7907 0.9302 0.4137      0
Review               0.9333      0.5000    0.6512 0.8837 0.4137      0
Viewpoint            0.0000      0.0000    0.0000 0.9612 0.4137      0
               Rmicro Pmicro
Case Study      18.67     28
Correspondence  18.67     28
Essay           18.67     28
Opinion         18.67     28
Perspective     18.67     28
Research        18.67     28
Review          18.67     28
Viewpoint       18.67     28

[[8]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.2857     0.2857   0.2857 0.9225 0.4844 0.4389
Correspondence        0.6667     0.8889   0.7619 0.9612 0.4844 0.4389
Essay                 0.4516     0.6667   0.5385 0.8140 0.4844 0.4389
Opinion               0.2500     0.9091   0.3922 0.7597 0.4844 0.4389
Perspective           0.3333     0.1364   0.1935 0.8062 0.4844 0.4389
Research              0.6667     0.6250   0.6452 0.9147 0.4844 0.4389
Review                0.8571     0.3636   0.5106 0.8217 0.4844 0.4389
Viewpoint             0.0000     0.0000   0.0000 0.9147 0.4844 0.4389
               Rmicro Pmicro
Case Study        5.9     59
Correspondence    5.9     59
Essay             5.9     59
Opinion           5.9     59
Perspective       5.9     59
Research          5.9     59
Review            5.9     59
Viewpoint         5.9     59

[[9]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.4444     0.4444   0.4444 0.9231 0.4681      0
Correspondence        0.5385     0.8750   0.6667 0.9462 0.4681      0
Essay                 0.3871     0.5714   0.4615 0.7846 0.4681      0
Opinion               0.3684     0.8750   0.5185 0.8000 0.4681      0
Perspective           0.0000     0.0000   0.0000 0.8769 0.4681      0
Research              0.7200     0.7200   0.7200 0.8923 0.4681      0
Review                0.6364     0.2593   0.3684 0.8154 0.4681      0
Viewpoint             0.0000     0.0000   0.0000 0.9154 0.4681      0
               Rmicro Pmicro
Case Study      5.636    Inf
Correspondence  5.636    Inf
Essay           5.636    Inf
Opinion         5.636    Inf
Perspective     5.636    Inf
Research        5.636    Inf
Review          5.636    Inf
```

```
Viewpoint       5.636    Inf

[[10]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5714    0.44444   0.5000 0.9375 0.5033  0.528
Correspondence        0.5789    0.91667   0.7097 0.9297 0.5033  0.528
Essay                 0.2973    0.64706   0.4074 0.7500 0.5033  0.528
Opinion               0.2432    0.90000   0.3830 0.7734 0.5033  0.528
Perspective           0.6667    0.09091   0.1600 0.8359 0.5033  0.528
Research              0.9375    0.83333   0.8824 0.9688 0.5033  0.528
Review                0.4286    0.10345   0.1667 0.7656 0.5033  0.528
Viewpoint             0.5000    0.09091   0.1538 0.9141 0.5033  0.528
               Rmicro Pmicro
Case Study      5.091     28
Correspondence  5.091     28
Essay           5.091     28
Opinion         5.091     28
Perspective     5.091     28
Research        5.091     28
Review          5.091     28
Viewpoint       5.091     28
```

Results of the SVM classifier per fold

```
lSVM = doCV(d2, 10, func_SVM)
lSVM
```

```
[[1]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.2500     0.1429   0.1818 0.9297 0.5505  0.584
Correspondence        0.9091     0.8333   0.8696 0.9766 0.5505  0.584
Essay                 0.6667     0.5000   0.5714 0.8828 0.5505  0.584
Opinion               0.7143     0.4545   0.5556 0.9375 0.5505  0.584
Perspective           0.5312     0.7727   0.6296 0.8438 0.5505  0.584
Research              0.7826     0.8571   0.8182 0.9375 0.5505  0.584
Review                0.8182     0.8438   0.8308 0.9141 0.5505  0.584
Viewpoint             0.0000     0.0000   0.0000 0.9531 0.5505  0.584
               Rmicro Pmicro
Case Study      29.33  29.33
Correspondence  29.33  29.33
Essay           29.33  29.33
Opinion         29.33  29.33
Perspective     29.33  29.33
Research        29.33  29.33
Review          29.33  29.33
Viewpoint       29.33  29.33

[[2]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.1429    0.07143  0.09524 0.8538 0.5214      0
Correspondence        1.0000    1.00000  1.00000 1.0000 0.5214      0
Essay                 0.8235    0.56000  0.66667 0.8923 0.5214      0
```

```
Opinion              0.1818    0.25000  0.21053 0.8846 0.5214      0
Perspective          0.4783    0.68750  0.56410 0.8692 0.5214      0
Research             0.5909    0.76471  0.66667 0.9000 0.5214      0
Review               0.7209    0.83784  0.77500 0.8615 0.5214      0
Viewpoint            0.0000    0.00000  0.00000 0.9538 0.5214      0
               Rmicro Pmicro
Case Study      13.17    Inf
Correspondence  13.17    Inf
Essay           13.17    Inf
Opinion         13.17    Inf
Perspective     13.17    Inf
Research        13.17    Inf
Review          13.17    Inf
Viewpoint       13.17    Inf

[[3]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5714     0.2667   0.3636 0.8915 0.5555      0
Correspondence        0.9167     0.9167   0.9167 0.9845 0.5555      0
Essay                 0.6154     0.5333   0.5714 0.9070 0.5555      0
Opinion               0.7143     0.3846   0.5000 0.9225 0.5555      0
Perspective           0.4062     0.7222   0.5200 0.8140 0.5555      0
Research              0.7273     0.7273   0.7273 0.9070 0.5555      0
Review                0.6944     0.8929   0.7812 0.8915 0.5555      0
Viewpoint             0.0000     0.0000   0.0000 0.9535 0.5555      0
               Rmicro Pmicro
Case Study      13.67    Inf
Correspondence  13.67    Inf
Essay           13.67    Inf
Opinion         13.67    Inf
Perspective     13.67    Inf
Research        13.67    Inf
Review          13.67    Inf
Viewpoint       13.67    Inf

[[4]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5000     0.1429   0.2222 0.8915   0.529      0
Correspondence        0.6667     1.0000   0.8000 0.9845   0.529      0
Essay                 0.5500     0.4400   0.4889 0.8217   0.529      0
Opinion               0.2500     0.2857   0.2667 0.9147   0.529      0
Perspective           0.5862     0.8095   0.6800 0.8760   0.529      0
Research              0.7083     0.7391   0.7234 0.8992   0.529      0
Review                0.5789     0.8148   0.6769 0.8372   0.529      0
Viewpoint             0.0000     0.0000   0.0000 0.9380   0.529      0
               Rmicro Pmicro
Case Study      9.375    Inf
Correspondence  9.375    Inf
Essay           9.375    Inf
Opinion         9.375    Inf
Perspective     9.375    Inf
Research        9.375    Inf
Review          9.375    Inf
Viewpoint       9.375    Inf
```

```
[[5]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study           1.00000     0.1000   0.1818 0.9302 0.5162      0
Correspondence       1.00000     0.6364   0.7778 0.9690 0.5162      0
Essay                0.60714     0.7727   0.6800 0.8760 0.5162      0
Opinion              0.09091     0.5000   0.1538 0.9147 0.5162      0
Perspective          0.37500     0.5000   0.4286 0.8140 0.5162      0
Research             0.79167     0.8261   0.8085 0.9302 0.5162      0
Review               0.79412     0.7941   0.7941 0.8915 0.5162      0
Viewpoint            0.00000     0.0000   0.0000 0.9302 0.5162      0
               Rmicro Pmicro
Case Study          9    Inf
Correspondence      9    Inf
Essay               9    Inf
Opinion             9    Inf
Perspective         9    Inf
Research            9    Inf
Review              9    Inf
Viewpoint           9    Inf

[[6]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            1.0000     0.3333   0.5000 0.9225 0.4883      0
Correspondence        0.8333     0.6667   0.7407 0.9457 0.4883      0
Essay                 0.4762     0.6667   0.5556 0.8760 0.4883      0
Opinion               0.2000     0.1429   0.1667 0.9225 0.4883      0
Perspective           0.3429     0.6000   0.4364 0.7597 0.4883      0
Research              0.6667     0.7273   0.6957 0.9457 0.4883      0
Review                0.7692     0.7692   0.7692 0.8605 0.4883      0
Viewpoint             0.0000     0.0000   0.0000 0.9457 0.4883      0
               Rmicro Pmicro
Case Study      10.86    Inf
Correspondence  10.86    Inf
Essay           10.86    Inf
Opinion         10.86    Inf
Perspective     10.86    Inf
Research        10.86    Inf
Review          10.86    Inf
Viewpoint       10.86    Inf

[[7]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5000     0.1111   0.1818 0.9302  0.491      0
Correspondence        0.8750     0.7000   0.7778 0.9690  0.491      0
Essay                 0.4091     0.4737   0.4390 0.8217  0.491      0
Opinion               0.2727     0.3750   0.3158 0.8992  0.491      0
Perspective           0.5000     0.5714   0.5333 0.7829  0.491      0
Research              0.8750     0.8750   0.8750 0.9535  0.491      0
Review                0.7667     0.8214   0.7931 0.9070  0.491      0
Viewpoint             0.0000     0.0000   0.0000 0.9767  0.491      0
               Rmicro Pmicro
Case Study      26.67    Inf
Correspondence  26.67    Inf
```

```
Essay           26.67    Inf
Opinion         26.67    Inf
Perspective     26.67    Inf
Research        26.67    Inf
Review          26.67    Inf
Viewpoint       26.67    Inf

[[8]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.1667     0.1429   0.1538 0.9147 0.5499      0
Correspondence        0.7000     0.7778   0.7368 0.9612 0.5499      0
Essay                 0.7222     0.6190   0.6667 0.8992 0.5499      0
Opinion               0.5000     0.4545   0.4762 0.9147 0.5499      0
Perspective           0.4167     0.6818   0.5172 0.7829 0.5499      0
Research              0.7368     0.8750   0.8000 0.9457 0.5499      0
Review                0.9333     0.8485   0.8889 0.9457 0.5499      0
Viewpoint             0.0000     0.0000   0.0000 0.9225 0.5499      0
               Rmicro Pmicro
Case Study        8.3    Inf
Correspondence    8.3    Inf
Essay             8.3    Inf
Opinion           8.3    Inf
Perspective       8.3    Inf
Research          8.3    Inf
Review            8.3    Inf
Viewpoint         8.3    Inf

[[9]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5000     0.2222   0.3077 0.9308 0.5221 0.6034
Correspondence        1.0000     0.6250   0.7692 0.9769 0.5221 0.6034
Essay                 0.7333     0.5238   0.6111 0.8923 0.5221 0.6034
Opinion               0.7500     0.1875   0.3000 0.8923 0.5221 0.6034
Perspective           0.2439     0.7692   0.3704 0.7385 0.5221 0.6034
Research              0.8000     0.9600   0.8727 0.9462 0.5221 0.6034
Review                0.8000     0.8889   0.8421 0.9308 0.5221 0.6034
Viewpoint             0.0000     0.0000   0.0000 0.9077 0.5221 0.6034
               Rmicro Pmicro
Case Study      7.182     79
Correspondence  7.182     79
Essay           7.182     79
Opinion         7.182     79
Perspective     7.182     79
Research        7.182     79
Review          7.182     79
Viewpoint       7.182     79

[[10]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            1.0000     0.3333   0.5000 0.9531 0.6051      0
Correspondence        1.0000     0.9167   0.9565 0.9922 0.6051      0
Essay                 0.5455     0.7059   0.6154 0.8828 0.6051      0
Opinion               0.4545     0.5000   0.4762 0.9141 0.6051      0
Perspective           0.5556     0.6818   0.6122 0.8516 0.6051      0
```

```
Research              0.7727      0.9444    0.8500 0.9531 0.6051       0
Review                0.6875      0.7586    0.7213 0.8672 0.6051       0
Viewpoint             0.0000      0.0000    0.0000 0.9141 0.6051       0
               Rmicro Pmicro
Case Study      7.727    Inf
Correspondence  7.727    Inf
Essay           7.727    Inf
Opinion         7.727    Inf
Perspective     7.727    Inf
Research        7.727    Inf
Review          7.727    Inf
Viewpoint       7.727    Inf
```

Results of the Random Forest classifier per fold

```
lRF = doCV(d2, 10, func_RF)
lRF
```

```
[[1]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.4000     0.5714   0.4706 0.9297  0.637 0.6227
Correspondence        0.8571     1.0000   0.9231 0.9844  0.637 0.6227
Essay                 0.7500     0.6000   0.6667 0.9062  0.637 0.6227
Opinion               0.5714     0.3636   0.4444 0.9219  0.637 0.6227
Perspective           0.6500     0.5909   0.6190 0.8750  0.637 0.6227
Research              0.7619     0.7619   0.7619 0.9219  0.637 0.6227
Review                0.8485     0.8750   0.8615 0.9297  0.637 0.6227
Viewpoint             0.1429     0.3333   0.2000 0.9375  0.637 0.6227
               Rmicro Pmicro
Case Study         30  12.86
Correspondence     30  12.86
Essay              30  12.86
Opinion            30  12.86
Perspective        30  12.86
Research           30  12.86
Review             30  12.86
Viewpoint          30  12.86

[[2]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.6154     0.5714   0.5926 0.9154 0.5461 0.5475
Correspondence        0.8571     0.8571   0.8571 0.9846 0.5461 0.5475
Essay                 0.8824     0.6000   0.7143 0.9077 0.5461 0.5475
Opinion               0.1667     0.1250   0.1429 0.9077 0.5461 0.5475
Perspective           0.3810     0.5000   0.4324 0.8385 0.5461 0.5475
Research              0.6087     0.8235   0.7000 0.9077 0.5461 0.5475
Review                0.8684     0.8919   0.8800 0.9308 0.5461 0.5475
Viewpoint             0.0000     0.0000   0.0000 0.9154 0.5461 0.5475
               Rmicro Pmicro
Case Study      14.17     17
Correspondence  14.17     17
Essay           14.17     17
```

```
Opinion          14.17     17
Perspective      14.17     17
Research         14.17     17
Review           14.17     17
Viewpoint        14.17     17

[[3]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study             0.4286     0.2000   0.2727 0.8760 0.5525 0.5584
Correspondence         0.9091     0.8333   0.8696 0.9767 0.5525 0.5584
Essay                  0.5238     0.7333   0.6111 0.8915 0.5525 0.5584
Opinion                0.4444     0.3077   0.3636 0.8915 0.5525 0.5584
Perspective            0.4783     0.6111   0.5366 0.8527 0.5525 0.5584
Research               0.6429     0.8182   0.7200 0.8915 0.5525 0.5584
Review                 0.8400     0.7500   0.7925 0.9147 0.5525 0.5584
Viewpoint              0.2000     0.1667   0.1818 0.9302 0.5525 0.5584
                Rmicro Pmicro
Case Study       13.17   15.8
Correspondence   13.17   15.8
Essay            13.17   15.8
Opinion          13.17   15.8
Perspective      13.17   15.8
Research         13.17   15.8
Review           13.17   15.8
Viewpoint        13.17   15.8

[[4]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study             0.6250     0.3571   0.4545 0.9070 0.5688 0.5611
Correspondence         0.6667     1.0000   0.8000 0.9845 0.5688 0.5611
Essay                  0.5789     0.4400   0.5000 0.8295 0.5688 0.5611
Opinion                0.3750     0.4286   0.4000 0.9302 0.5688 0.5611
Perspective            0.5000     0.5714   0.5333 0.8372 0.5688 0.5611
Research               0.8095     0.7391   0.7727 0.9225 0.5688 0.5611
Review                 0.6000     0.8889   0.7164 0.8527 0.5688 0.5611
Viewpoint              0.3333     0.1250   0.1818 0.9302 0.5688 0.5611
                Rmicro Pmicro
Case Study       9.625  25.67
Correspondence   9.625  25.67
Essay            9.625  25.67
Opinion          9.625  25.67
Perspective      9.625  25.67
Research         9.625  25.67
Review           9.625  25.67
Viewpoint        9.625  25.67

[[5]]
                tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.50000     0.4000   0.4444 0.9225 0.5328 0.5255
Correspondence        1.00000     0.7273   0.8421 0.9767 0.5328 0.5255
Essay                 0.65217     0.6818   0.6667 0.8837 0.5328 0.5255
Opinion               0.08333     0.5000   0.1429 0.9070 0.5328 0.5255
Perspective           0.33333     0.3333   0.3333 0.8140 0.5328 0.5255
Research              0.86364     0.8261   0.8444 0.9457 0.5328 0.5255
```

```
Review               0.77143      0.7941    0.7826 0.8837 0.5328 0.5255
Viewpoint            0.00000      0.0000    0.0000 0.9070 0.5328 0.5255
               Rmicro Pmicro
Case Study      8.889  26.67
Correspondence  8.889  26.67
Essay           8.889  26.67
Opinion         8.889  26.67
Perspective     8.889  26.67
Research        8.889  26.67
Review          8.889  26.67
Viewpoint       8.889  26.67

[[6]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.8000     0.2667   0.4000 0.9070  0.526 0.5711
Correspondence        0.7368     0.9333   0.8235 0.9535  0.526 0.5711
Essay                 0.4737     0.6000   0.5294 0.8760  0.526 0.5711
Opinion               0.2500     0.1429   0.1818 0.9302  0.526 0.5711
Perspective           0.4615     0.6000   0.5217 0.8295  0.526 0.5711
Research              0.5714     0.7273   0.6400 0.9302  0.526 0.5711
Review                0.7750     0.7949   0.7848 0.8682  0.526 0.5711
Viewpoint             0.5000     0.1429   0.2222 0.9457  0.526 0.5711
               Rmicro Pmicro
Case Study      11.43     40
Correspondence  11.43     40
Essay           11.43     40
Opinion         11.43     40
Perspective     11.43     40
Research        11.43     40
Review          11.43     40
Viewpoint       11.43     40

[[7]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            1.0000     0.4444   0.6154 0.9612 0.5991 0.6652
Correspondence        0.8000     0.8000   0.8000 0.9690 0.5991 0.6652
Essay                 0.5455     0.6316   0.5854 0.8682 0.5991 0.6652
Opinion               0.6667     0.5000   0.5714 0.9535 0.5991 0.6652
Perspective           0.6429     0.6429   0.6429 0.8450 0.5991 0.6652
Research              0.9167     0.9167   0.9167 0.9690 0.5991 0.6652
Review                0.7500     0.8571   0.8000 0.9070 0.5991 0.6652
Viewpoint             0.0000     0.0000   0.0000 0.9535 0.5991 0.6652
               Rmicro Pmicro
Case Study      30.67  30.67
Correspondence  30.67  30.67
Essay           30.67  30.67
Opinion         30.67  30.67
Perspective     30.67  30.67
Research        30.67  30.67
Review          30.67  30.67
Viewpoint       30.67  30.67

[[8]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
```

```
Case Study            0.2222     0.2857   0.2500 0.9070 0.5595 0.5626
Correspondence        0.7778     0.7778   0.7778 0.9690 0.5595 0.5626
Essay                 0.6364     0.6667   0.6512 0.8837 0.5595 0.5626
Opinion               0.5455     0.5455   0.5455 0.9225 0.5595 0.5626
Perspective           0.4074     0.5000   0.4490 0.7907 0.5595 0.5626
Research              0.6500     0.8125   0.7222 0.9225 0.5595 0.5626
Review                0.9286     0.7879   0.8525 0.9302 0.5595 0.5626
Viewpoint             0.3333     0.1000   0.1538 0.9147 0.5595 0.5626
               Rmicro Pmicro
Case Study          8  26.67
Correspondence      8  26.67
Essay               8  26.67
Opinion             8  26.67
Perspective         8  26.67
Research            8  26.67
Review              8  26.67
Viewpoint           8  26.67

[[9]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5000    0.33333   0.4000 0.9308 0.5314 0.5526
Correspondence        0.7143    0.62500   0.6667 0.9615 0.5314 0.5526
Essay                 0.6000    0.57143   0.5854 0.8692 0.5314 0.5526
Opinion               0.5000    0.43750   0.4667 0.8769 0.5314 0.5526
Perspective           0.2609    0.46154   0.3333 0.8154 0.5314 0.5526
Research              0.7586    0.88000   0.8148 0.9231 0.5314 0.5526
Review                0.9200    0.85185   0.8846 0.9538 0.5314 0.5526
Viewpoint             0.1667    0.09091   0.1176 0.8846 0.5314 0.5526
               Rmicro Pmicro
Case Study      7.182  13.17
Correspondence  7.182  13.17
Essay           7.182  13.17
Opinion         7.182  13.17
Perspective     7.182  13.17
Research        7.182  13.17
Review          7.182  13.17
Viewpoint       7.182  13.17

[[10]]
               tab.precision tab.recall fmeasure    acc Rmacro Pmacro
Case Study            0.5000    0.33333   0.4000 0.9297 0.5671 0.5661
Correspondence        0.9167    0.91667   0.9167 0.9844 0.5671 0.5661
Essay                 0.6000    0.70588   0.6486 0.8984 0.5671 0.5661
Opinion               0.2105    0.40000   0.2759 0.8359 0.5671 0.5661
Perspective           0.5385    0.31818   0.4000 0.8359 0.5671 0.5661
Research              0.7391    0.94444   0.8293 0.9453 0.5671 0.5661
Review                0.7742    0.82759   0.8000 0.9062 0.5671 0.5661
Viewpoint             0.2500    0.09091   0.1333 0.8984 0.5671 0.5661
               Rmicro Pmicro
Case Study      7.182  19.75
Correspondence  7.182  19.75
Essay           7.182  19.75
Opinion         7.182  19.75
Perspective     7.182  19.75
```

```
Research         7.182  19.75
Review           7.182  19.75
Viewpoint        7.182  19.75
```

Let's now compute the average of each fold for each measures.

```
#get the average accuracy for each fold and each classifier
getAverage = function(l){
  folds <- c('Fold 1', 'Fold 2', 'Fold 3', 'Fold 4', 'Fold 5', 'Fold 6', '
     Fold 7', 'Fold 8', 'Fold 9', 'Fold 10', 'Average')
  x <- numeric(11)
  df <- data.frame(x,x,x,x,x,x,x,x, row.names=folds)
  colnames(df) <- c('Precision', 'Recall', 'Fmeasure', 'Accuracy', 'Rmacro
     ', 'Pmacro', 'Rmicro', 'Pmicro')
  #browse by column
  for(j in 1:dim(l[[1]])[1]){
    #browse by folds
    for(i in 1:length(l)){
      df[i,j] <- sapply(l[[i]][j], mean)
    }
  }
  df[11,] <- apply(df[-11,], 2, mean)
  return(df)
}
```

Average of the Naive Bayes classifier per fold

```
dfNaive <- getAverage(lNaive)
dfNaive
```

```
         Precision Recall Fmeasure Accuracy Rmacro Pmacro Rmicro
Fold 1      0.4347 0.5170   0.4288   0.8750 0.5170 0.0000 21.333
Fold 2      0.5093 0.4755   0.4395   0.8615 0.4755 0.5093  9.667
Fold 3      0.4309 0.4715   0.4134   0.8721 0.4715 0.4309 10.500
Fold 4      0.4481 0.4142   0.3190   0.8372 0.4142 0.4481  5.625
Fold 5      0.3926 0.4702   0.3577   0.8469 0.4702 0.3926  5.556
Fold 6      0.3357 0.3642   0.3066   0.8391 0.3642 0.3357  6.571
Fold 7      0.3962 0.4137   0.3678   0.8585 0.4137 0.0000 18.667
Fold 8      0.4389 0.4844   0.4159   0.8643 0.4844 0.4389  5.900
Fold 9      0.3868 0.4681   0.3974   0.8692 0.4681 0.0000  5.636
Fold 10     0.5280 0.5033   0.4204   0.8594 0.5033 0.5280  5.091
Average     0.4301 0.4582   0.3867   0.8583 0.4582 0.3084  9.455
         Pmicro
Fold 1    32.00
Fold 2    29.00
Fold 3    63.00
Fold 4    45.00
Fold 5    16.67
Fold 6    23.00
Fold 7    28.00
Fold 8    59.00
Fold 9      Inf
```

```
Fold 10  28.00
Average    Inf
```

Average of the SVM classifier per fold

```
dfSVM <- getAverage(lSVM)
dfSVM
```

```
         Precision Recall Fmeasure Accuracy Rmacro Pmacro Rmicro
Fold 1      0.5840 0.5505   0.5571   0.9219 0.5505 0.5840 29.333
Fold 2      0.4923 0.5214   0.4973   0.9019 0.5214 0.0000 13.167
Fold 3      0.5807 0.5555   0.5475   0.9089 0.5555 0.0000 13.667
Fold 4      0.4800 0.5290   0.4823   0.8953 0.5290 0.0000  9.375
Fold 5      0.5824 0.5162   0.4781   0.9070 0.5162 0.0000  9.000
Fold 6      0.5360 0.4883   0.4830   0.8973 0.4883 0.0000 10.857
Fold 7      0.5248 0.4910   0.4895   0.9050 0.4910 0.0000 26.667
Fold 8      0.5220 0.5499   0.5300   0.9109 0.5499 0.0000  8.300
Fold 9      0.6034 0.5221   0.5092   0.9019 0.5221 0.6034  7.182
Fold 10     0.6270 0.6051   0.5915   0.9160 0.6051 0.0000  7.727
Average     0.5533 0.5329   0.5165   0.9066 0.5329 0.1187 13.527
         Pmicro
Fold 1   29.33
Fold 2     Inf
Fold 3     Inf
Fold 4     Inf
Fold 5     Inf
Fold 6     Inf
Fold 7     Inf
Fold 8     Inf
Fold 9   79.00
Fold 10    Inf
Average    Inf
```

Average of the Random Forest classifier per fold

```
dfRF <- getAverage(lRF)
dfRF
```

```
         Precision Recall Fmeasure Accuracy Rmacro Pmacro Rmicro
Fold 1      0.6227 0.6370   0.6184   0.9258 0.6370 0.6227 30.000
Fold 2      0.5475 0.5461   0.5399   0.9135 0.5461 0.5475 14.167
Fold 3      0.5584 0.5525   0.5435   0.9031 0.5525 0.5584 13.167
Fold 4      0.5611 0.5688   0.5449   0.8992 0.5688 0.5611  9.625
Fold 5      0.5255 0.5328   0.5071   0.9050 0.5328 0.5255  8.889
Fold 6      0.5711 0.5260   0.5129   0.9050 0.5260 0.5711 11.429
Fold 7      0.6652 0.5991   0.6165   0.9283 0.5991 0.6652 30.667
Fold 8      0.5626 0.5595   0.5502   0.9050 0.5595 0.5626  8.000
Fold 9      0.5526 0.5314   0.5336   0.9019 0.5314 0.5526  7.182
Fold 10     0.5661 0.5671   0.5505   0.9043 0.5671 0.5661  7.182
Average     0.5733 0.5620   0.5517   0.9091 0.5620 0.5733 14.031
```

```
         Pmicro
Fold 1    12.86
Fold 2    17.00
Fold 3    15.80
Fold 4    25.67
Fold 5    26.67
Fold 6    40.00
Fold 7    30.67
Fold 8    26.67
Fold 9    13.17
Fold 10   19.75
Average   22.82
```

We can observe that the Naïve Bayes classifier reaches an average accuracy of 85%. The SVM classifier reaches 90% of accuracy and the Random Forest one is around 91%. Therefore, the SVM and Random Forest classifiers seem to perform better than the Naïve Bayes one.

```
getTvalue = function(dfNaive, dfSVM, dfRF){
  folds <- c('Fold 1', 'Fold 2', 'Fold 3', 'Fold 4', 'Fold 5', 'Fold 6', '
     Fold 7', 'Fold 8', 'Fold 9', 'Fold 10', 'Average', 'Stdev', 't-value
     ')
  cols <- c('NB-SVM', 'NB-RF', 'SVM-RF')
  x <- numeric(13)
  df <- data.frame(x,x,x, row.names=folds)
  #extract the average and standard deviation of each of the three
     classfiers to calculate the t-test
  for(i in 1:nrow(dfNaive)){
    df[i,1] <- dfNaive[i,4] - dfSVM[i,4]
    df[i,2] <- dfNaive[i,4] - dfRF[i,4]
    df[i,3] <- dfSVM[i,4] - dfRF[i,4]
  }
  df[11,] <- apply(df[-c(11,12,13),], 2, mean)
  df[12,] <- apply(df[-c(11,12,13),], 2, sd)
  for(i in 1:ncol(df)){
    df[13,i] <- abs(df[11,i]/(df[12,i]/sqrt(10)))
  }
  colnames(df) <- cols
  return (df)
}
dTvalue <- getTvalue(dfNaive, dfSVM, dfRF)
dTvalue
```

```
           NB-SVM    NB-RF    SVM-RF
Fold 1   -0.046875 -0.05078 -0.003906
Fold 2   -0.040385 -0.05192 -0.011538
Fold 3   -0.036822 -0.03101  0.005814
Fold 4   -0.058140 -0.06202 -0.003876
Fold 5   -0.060078 -0.05814  0.001938
Fold 6   -0.058140 -0.06589 -0.007752
Fold 7   -0.046512 -0.06977 -0.023256
Fold 8   -0.046512 -0.04070  0.005814
Fold 9   -0.032692 -0.03269  0.000000
Fold 10  -0.056641 -0.04492  0.011719
Average  -0.048279 -0.05078 -0.002504
```

```
Stdev     0.009709   0.01344   0.010057
t-value  15.725308  11.95267   0.787455
```

As we can observe, the first two columns have a very high t-test value. Only the third column has a t-test value which we can compare with the t-student's table. These results are highly related with the results concerning the accuracy.

- We showed earlier that the SVM classifier was 5% more accurate than the Naive Bayes classifier. We can assume that the SVM classifier performs better than the Naive Bayes one because the t-test reveals that we are 99% confident we can reject the null hypothesis that these two algorithms perform the same. This 99% confidence comes from the fact that the t-test value is equal to 15.725308. This value cannot be compared to any other values within the t-student's table.
- Let's now compare the Naive Bayes and Random Forest classifiers. We can conclude exactly the same conclusion than the previous one. We are 99% confident that Random Forest classifier performs better than the Naive Bayes one.
- Concerning the last comparison between the SVM and Random Forest algorithms, SVM performed with 90.66% and the Random Forest performed with 90.91%. So far, Random Forest seems a more accurate classifier in this case. Refering to the t-student's table, this t-test value (0.787455) shows that we are between 75 and 80% confident that the Random Forest algorithm performs better as shows the accuracy. Therefore, we are between 20 and 25% not confident about this hypothesis.

As a conclusion, the Random Forest classifier seems to perform better than the other classifiers.