Project 2 - Spam Filter

Rohman Sultan

November 8, 2021

1 Introduction

This project is about training and classifying messages to determine whether it is a spam or ham. The objective is the user feeds the program with data set that the program to train on. The program will then save the trained model to a file to be used later for classifying new data.

2 Problems

For this project I have decided to use R because after some research, it's great for statistical computing and that aligns perfectly with the project's objective, as well as another opportunity for me to learn a new language. However, I found R to be more difficult to use because I never used it before and it's more of a functional language, which I am not used to. The major problem I was facing was the extremely slow running time of my program when training and classifying the data. R is not known to be for speed but conveniences for the user. Writing efficient code is very important for large data sets. Another problem was how the data should be interpreted and processed by R. The CVS file had problems in it such as miss placed commas, strange characters, and symbols. This required manual removal and cleanup before proceeding. The contents of the file is actually a good representative of how a real world data set might look like.

3 Approach

To lower the long computing time cost, the training will work on a subset of the data. The program will process just the words with each row without considering numbers, punctuation, and stop words because they provide little value when classifying. The two output files specified by the user will contain the total number of words and word frequency for ham and spam when running the training script. The program will not actually reuse the output files when classifying but will instead reuse a model for better efficiency.

4 File structure

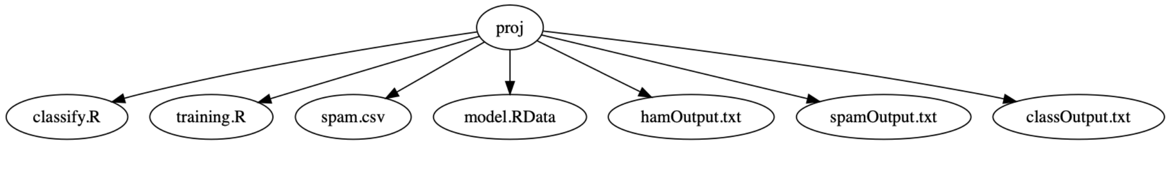
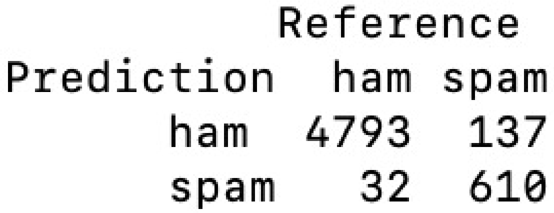


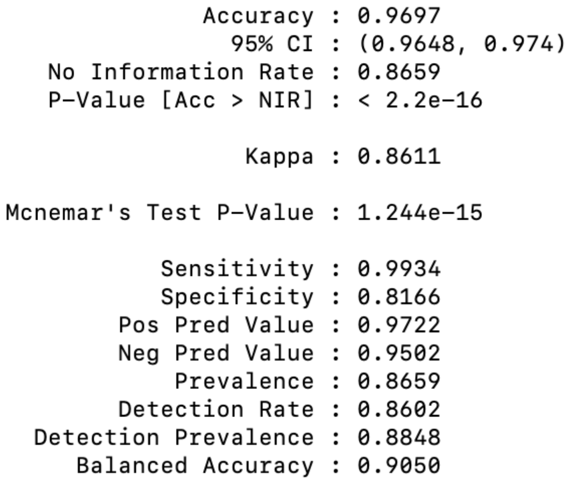
Figure 1: Project 2 file structure

5 Results and discussions



```
              Reference
Prediction  ham spam
      ham  4793  137
      spam   32  610
```

Figure 2: Confusion matrix



```
               Accuracy : 0.9697
                 95% CI : (0.9648, 0.974)
    No Information Rate : 0.8659
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8611

 Mcnemar's Test P-Value : 1.244e-15

            Sensitivity : 0.9934
            Specificity : 0.8166
         Pos Pred Value : 0.9722
         Neg Pred Value : 0.9502
             Prevalence : 0.8659
         Detection Rate : 0.8602
   Detection Prevalence : 0.8848
      Balanced Accuracy : 0.9050
```

Figure 3: Results obtained

I was able to achieve an accuracy of 97% with the supplied data set. The model only mislabeled 32 spam messages as ham. Overall, the model is an adequate spam classifier.