Configuration Manual

MSc Research Project
Data Analytics

Princy Dcunha
Student ID: x18135889

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Princy Dcunha
Student ID:	x18135889
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	12/12/2019
Project Title:	Configuration Manual
Word Count:	903
Page Count:	9

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Princy Dcunha
x18135889

Introduction

This Configuration Manual aims at illustrating all the steps done to get to the end of this project. Right from stating the hardware 1 used to the various softwares 2 and tools used, is mentioned in the next pages. The outputs and R Scripts that could not be put into the technical report are attached below as well.

1 Hardware Configuration

1.1 MacBook Pro, 2018

The Figure 1 shows the Mac OS configuration used for all processes of this project. It is updated to the latest version 10.14.6 (18G1012) MacOS Mojave and has a 2.2 GHz Intel Core i7 processor.

Figure 1: MAC OS

2 Software Configuration

2.1 RStudio

Figure 2 shows the RStudio version used for running all the R scripts from cleaning and preparing the data to implementing and evaluating the models. The version used was RStudio Desktop 1.2.5019. Figure 3 Shows the successful installation of RStudio Software.

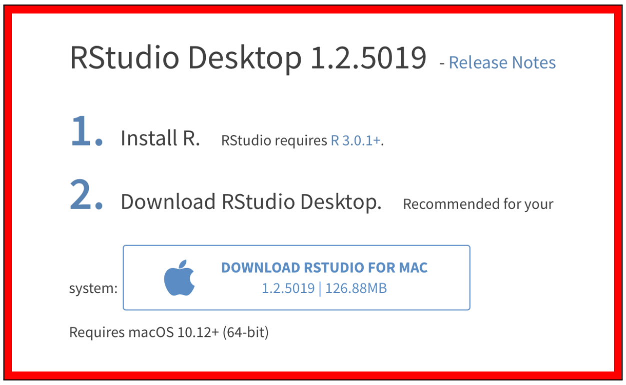

Figure 2: RStudio

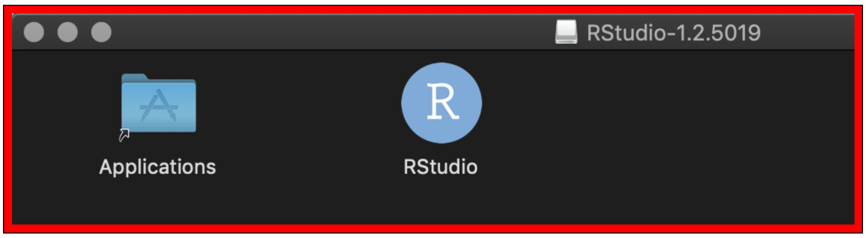

Figure 3: RStudio Installed

2.2 Overleaf

This Figure 4 online documentation tool was used for all the documentations related to this project. It has an inbuilt library and is like an html code with tags and labels. Everything done on overleaf is automatically saved on the cloud, eliminating the risk of losing a drafted document.

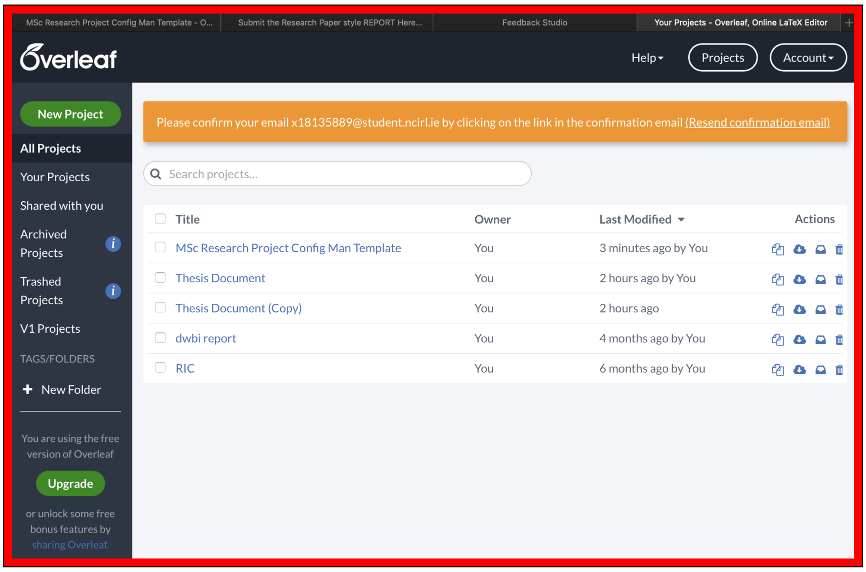

Figure 4: Overleaf

2.3 Microsoft Excel

Figure 5 displays the version of Microsoft Excel for Mac used , which is Version 16.30 . Figure 6 shows an excel sheet containing the project dataset . Microsoft Excel was used to do minor tweaks and adjustments and VLOOKUP() was used to combine two sheets having common columns.

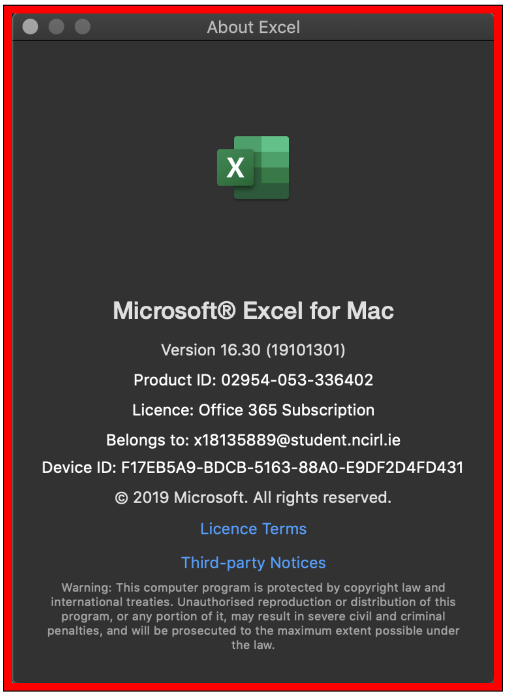

Figure 5: Microsoft Excel Version

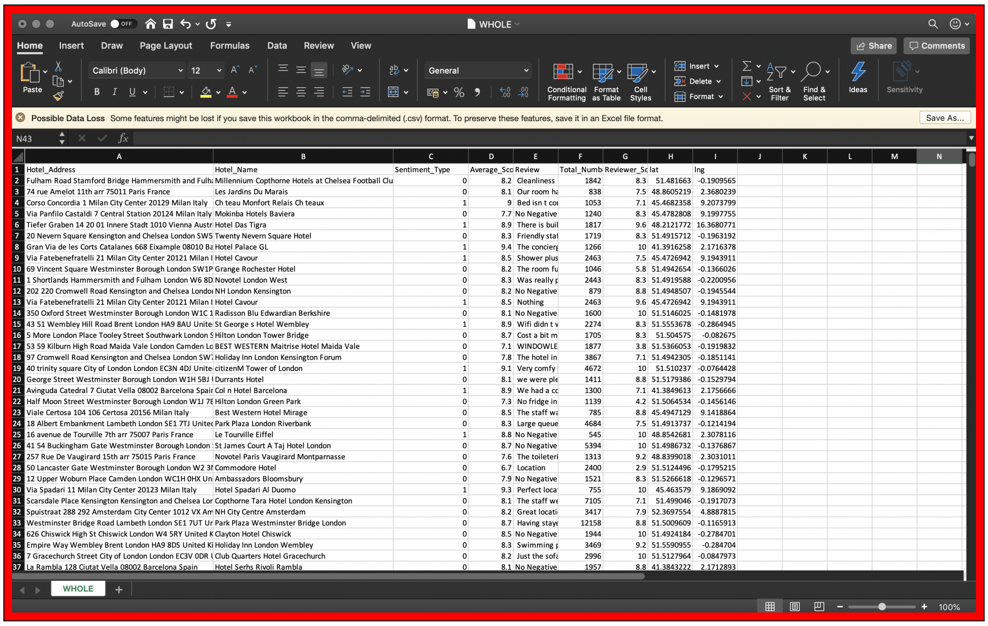

Figure 6: Microsoft Excel Sheet

2.4 Web

The web plays an important role in supporting the completion of any project. Numerous amount of data and knowledge was gained form the web for stating and implementing this project.

2.4.1 Chrome

The Google Chrome Figure 7 was used for finding datasets and to explore various possibilities to complete and complement the project.

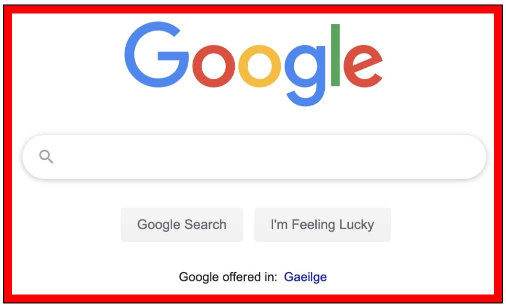

Figure 7: Google

2.4.2 Safari

The Safari Figure 8 is a web browser made for MacOS and comes in very handy when in need. It was the default browser used for most of the job.

Figure 8: Safari

2.5 Outputs and Visualizations

Figure 9 shows the distribution of reviewer scores in the dataset. This was a part of exploratory data analysis.

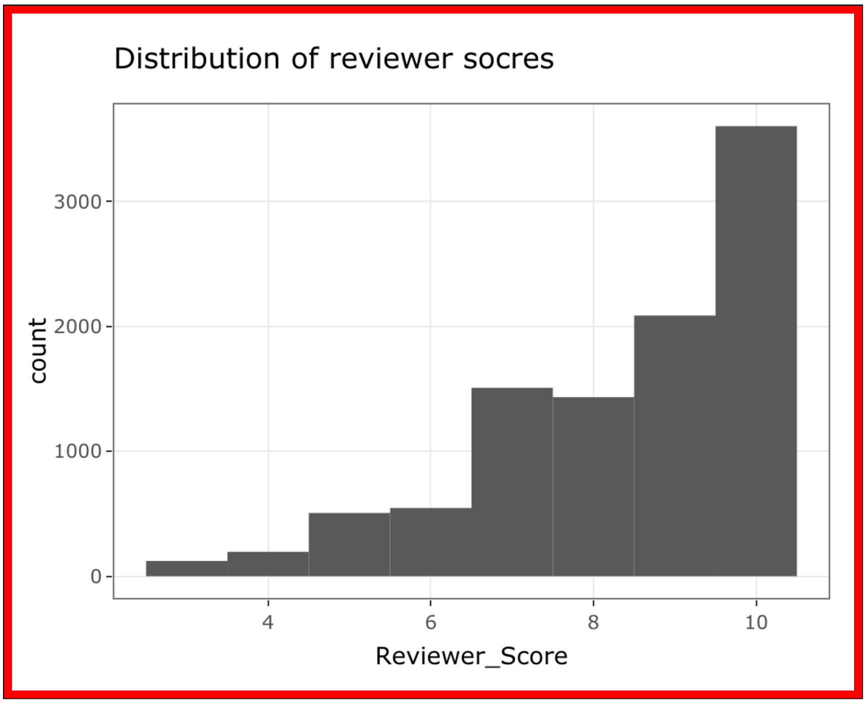

Figure 9: Reviewer_Score Plot

Figure 10 plots the Average Reviewer Score from the dataset. This also was a part of Exploratory Data Analysis (EDA).

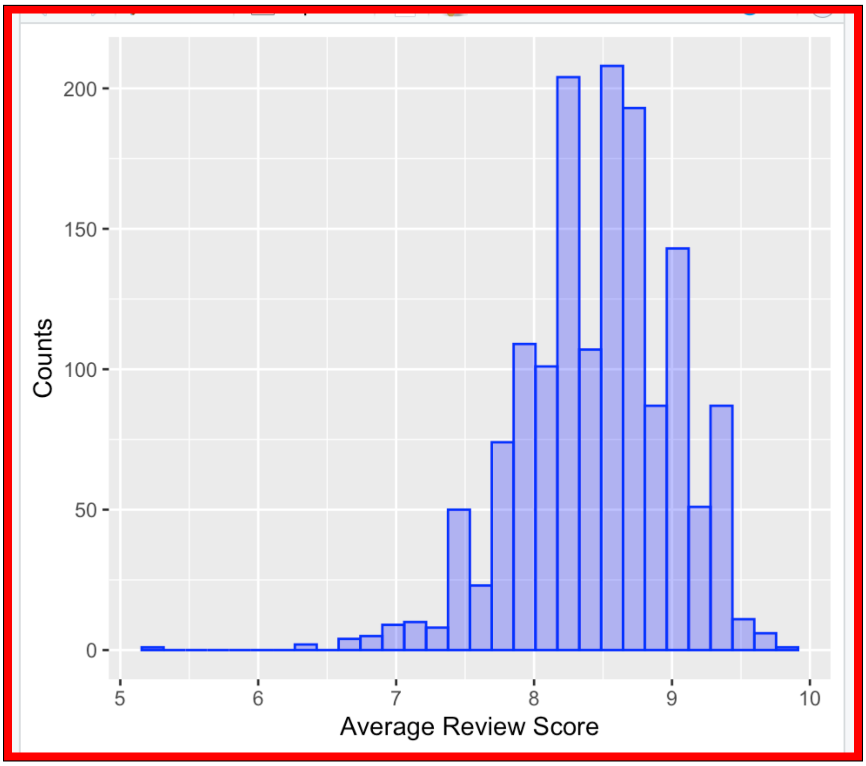

Figure 10: Average Review Score

Figure 11 shows the Top most score of the ratings and Figure 12 shows the bottom most score of the user ratings.

```
> data.frame(percentile=perc,score=score)->d
> print("Top rating scores are:")
[1] "Top rating scores are:"
> d%>%arrange(desc(score))%>%head(5)
  percentile score
1        98%   9.4
2        99%   9.4
3        96%   9.3
4        97%   9.3
5        93%   9.2
```

Figure 11: Top Reviews

```
> print("Bottom rating scores are:")
[1] "Bottom rating scores are:"
> d%>%arrange(score)%>%head(5)
  percentile score
1         1%   6.9
2         2%   7.1
3         3%   7.1
4         4%   7.3
5         5%   7.4
>
```

Figure 12: Bottom Reviews

Figure 13 displays the output of topic modelling. It is not clear because it does not have unique topics, rather it has unique full sentences, which makes it difficult to plot as per the topic. This was not a very successful approach and would be recommended as future work.

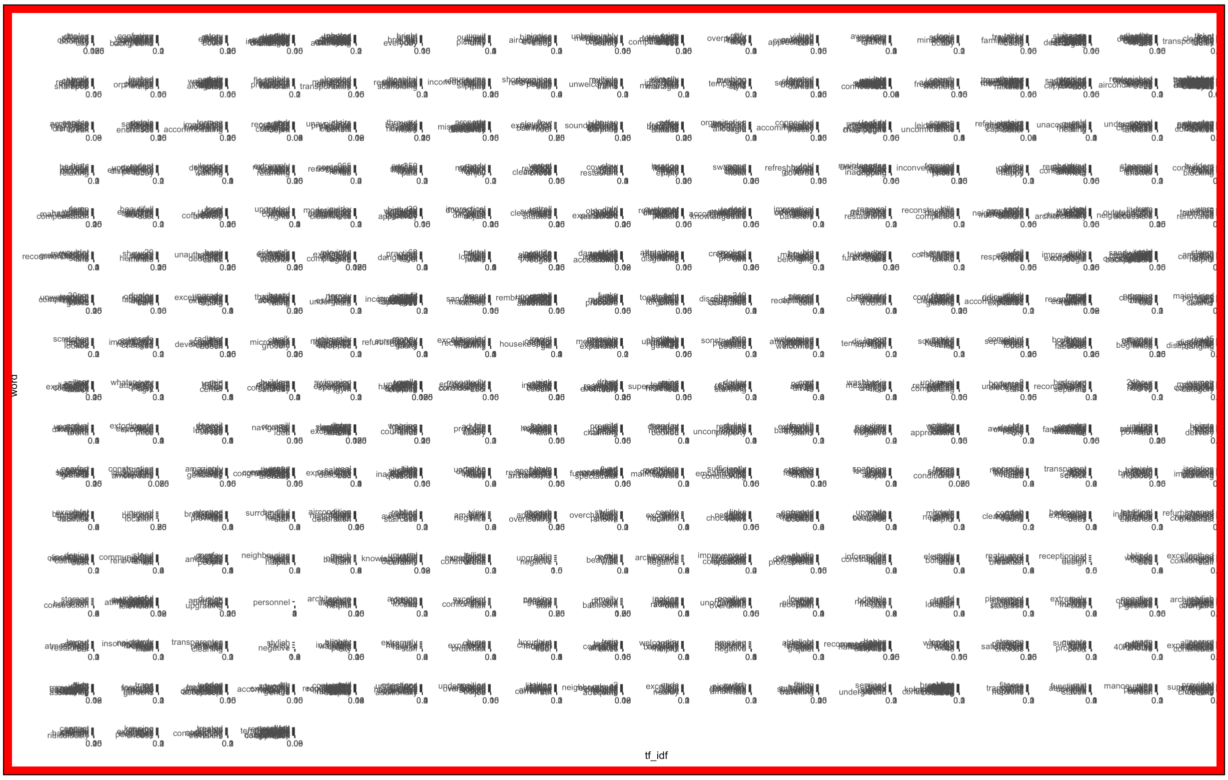

Figure 13: Topic Modelling

Figure 14 shows the extraction of countries from the hotel address.

```
> countries=paste(unique(df$country.x),collapse=",")
> message=paste("The countries mentioned in the dataset are:", countries)
> print(message)
[1] "The countries mentioned in the dataset are: Netherlands,United Kingdom,France,Spain,Italy,Austria"
>
```

Figure 14: Countries mentioned in dataset

Figure 15 show the extraction of cities from the address and country names.

```
> cities=paste(unique(df$city.x),collapse=",")
> message=paste("The cities mentioned in the dataset are:",  cities)
> print(message)
[1] "The cities mentioned in the dataset are: Amsterdam,London,Paris,Barcelona,Milan,Vienna"
>
```

Figure 15: Cities in dataset

Figure 16 gives the bar plot of aspects extracted from topic modelling, showing its count and relevance in that specific review/sentence. Its the Term Frequency-Inverse Document Frequency (TF-IDF) in which we can see how important a word was to that sentence.

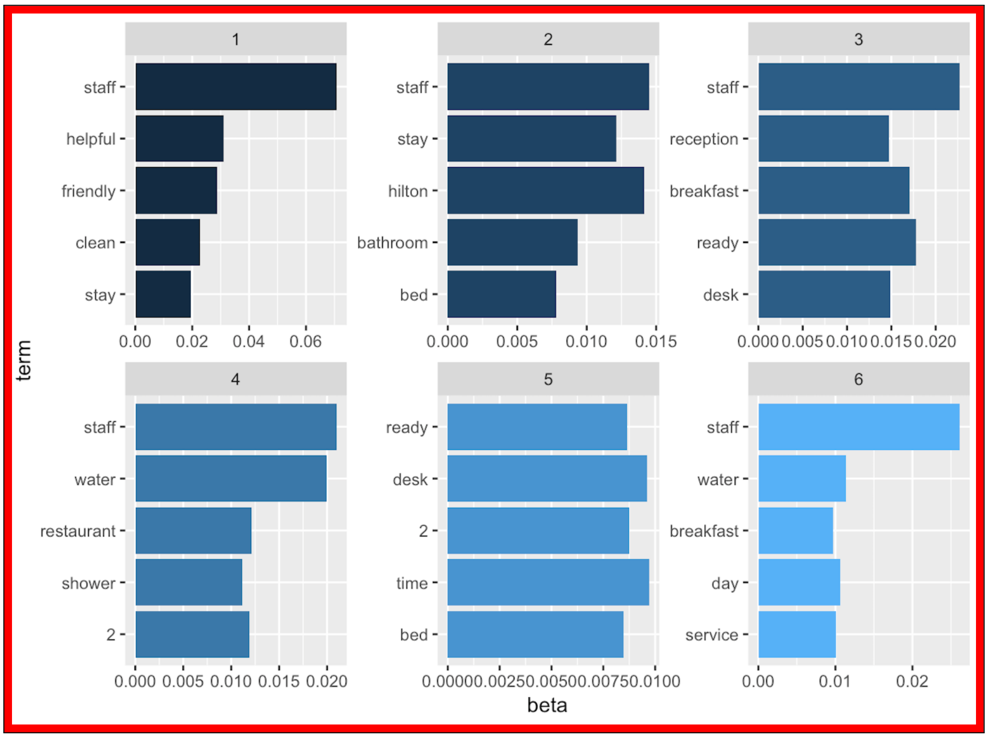

Figure 16: TF-IDF

Figure 17 shows how much an aspect has an effect on the given sentence.

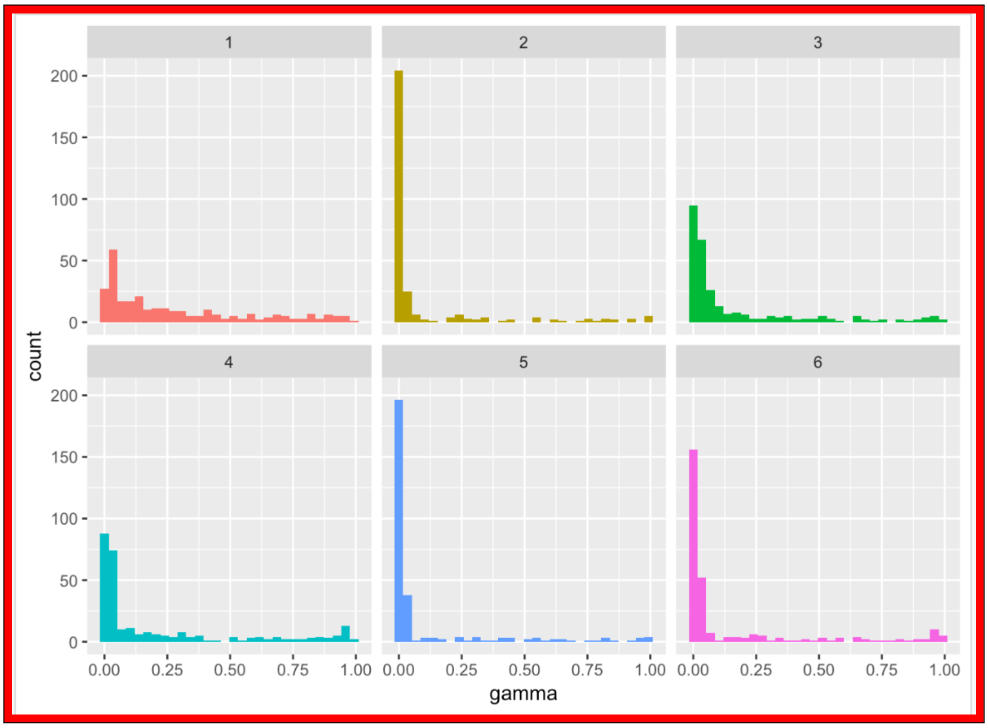

Figure 17: Topic Modelling

Figure 18 shows how the data is split into trainig set and test set.

```
dataset = read.csv("/Users/princydcunha/Desktop/NewWhole.csv")
#View(dataset)
str(dataset)
dataset = dataset[4:8]
#dataset$Sentiment_Type= factor(dataset$Sentiment_Type, levels = c(0, 1))
dataset$Review = NULL
dataset$Sentiment_Type = as.numeric(dataset$Sentiment_Type)
dataset$Average_Score = as.numeric(dataset$Average_Score )
dataset$Total_Number_of_Reviews = as.numeric(dataset$Total_Number_of_Reviews)
dataset$Reviewer_Score= as.numeric(dataset$Reviewer_Score)


set.seed(123)
split = sample.split(dataset$Sentiment_Type, SplitRatio = 0.75)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
#View(training_set)
#View(test_set)

#Feature scaling
training_set[-1] = scale(training_set[-1])
test_set[-1] = scale(test_set[-1])

training_set$Sentiment_Type = as.factor(training_set$Sentiment_Type)
test_set$Sentiment_Type = as.factor(test_set$Sentiment_Type)
```

Figure 18: Splitting of Train and Test Data

Figure 19 shows the code used to plot the confusion matrix in Random Forest algortihm.

```
#=============   Random Forest


model1 <- randomForest(Sentiment_Type ~ ., data = training_set, importance = TRUE)
model1
varImpPlot(model1)

confusionMatrix(predict(model1,test_set), test_set$Sentiment_Type)
```

Figure 19:

Figure 20 shows the ANOVA results using ChiSquare Test.

```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Sentiment_Type

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                     6999     9284.1
Average_Score            1   3217.1      6998     6067.0   <2e-16 ***
Total_Number_of_Reviews  1    363.4      6997     5703.6   <2e-16 ***
Reviewer_Score           1      1.8      6996     5701.8   0.1744
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20:

Figure 21 displays the R code used to calculate and predict the accuracy in Logistic Regression.

```
pR2(model)
fitted.results <- predict(model,newdata=subset(test,select=c(2,3,4)),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$Sentiment_Type)
print(paste('Accuracy',1-misClasificError))    # "Accuracy 0.821631878557875"

p <- predict(model, newdata=subset(test,select=c(2,3,4)), type="response")
pr <- prediction(p, test$Sentiment_Type)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]  #auc 0.8875061
```

Figure 21: