

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.

<u>ALL</u> 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.

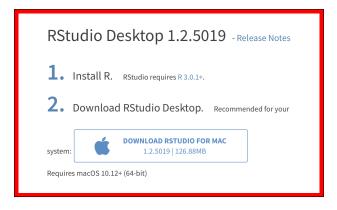


Figure 2: RStudio

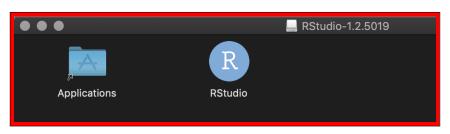


Figure 3: RStudio Installed

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.

MSc Research Project Config	g Man Template - O Submit the Research Paper style REPORT Here	. Feedback Stuc	Help • Projects	af, Online LaTeX Editor
New Project	Please confirm your email x18135889@student.nci	rl.ie by clicking on the link i	n the confirmation email <u>(Resend cc</u>	onfirmation email)
All Projects				
Your Projects	Q Search projects			
Shared with you	□ Title	Owner	Last Modified 💌	Actions
Archived Projects	MSc Research Project Config Man Template	You	3 minutes ago by You	4 0 0 f
Trashed	Thesis Document	You	2 hours ago by You	42 44 44 f
Projects 🔍	Thesis Document (Copy)	You	2 hours ago	20 00 0 t
V1 Projects	dwbi report	You	4 months ago by You	20 00 00 t
TAGS/FOLDERS		You	6 months ago by You	2 4 4 1
+ New Folder				
You are using the free version of Overleaf				
Upgrade				
or unlock some free bonus features by sharing Overleaf.				

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

🔍 🌢 🛛 AutoSave 🌒 📭 🏠 🕤	v O ÷				WHOL	.E ~									
me Insert Draw Page Layout	Formulas	Data Review	View										ල් Sł	iare 📮	Comme
A _ Å Calibri (Body) → 1	2 v A A		89 • I	🥸 🗸 🛛 Genera			u • 🖩	B. 188	. 💷 i	nsert v	Σ • Ας	7 • O • t& Find & ter Select			
							🖿 , fi	ø• ≞ ∕/	້ 🖼 ເ	Delete 🗸		r ×·	7		
aste 🦽 🛛 B I 🛛 🗸 🖾 🗸 🗌	🙅 🗸 🖌 🗸		€ 30	🗄 v 🛛 🖙 v 😨		8 38 60	nditional For	mat Cell	. mini	format v	V Sor	t& Find &	Ideas		
▼ 1						1 PO	maxing as i	lable Styles	· · · · · ·	ormat - 1	A : 10	tel Select	1		
Possible Data Loss Some features might be	e lost if you save th	nis workbook in the c	omma-delimit	led (.csv) format. 1	o preserve th	ese features, :	save it in an E	Excel file form	nat.						Save A
$f_X \to f_X$															
A		в		c	D	ε	F	G	н	1	ر ا	к	ι	м	N
lotel Address	Hotel Name			Sentiment Type	Average S	co Beview	Total Numb	Reviewer Scl	at	Ing					
ulham Road Stamford Bridge Hammersmith and		othorne Hotels at Chel	sea Football Cl			2 Cleanliness	1842			-0.1909565					
4 rue Amelot 11th arr 75011 Paris France	Les Jardins Du	Marais			0 8	1 Our room h	838	7.5	48.8605219	2.3680239					
Corso Concordia 1 Milan City Center 20129 Milan II	taly Chiteau Monfo	rt Relais Ch teaux			1	9 Bed isn t co	1053	7.1	45.4682358	9.2073799					
/ia Panfilo Castaldi 7 Central Station 20124 Milan	Italy Mokinba Hotel	s Baviera			0 7	7 No Negative	1240	8.3	45.4782808	9.1997755					
iefer Graben 14 20 01 Innere Stadt 1010 Vienna	Austr Hotel Das Tigr	a			1 8	.9 There is bui	1817	9.6	48.2121772	16.3680771					
0 Nevern Square Kensington and Chelsea London	SWS Twenty Neven	n Square Hotel			0 8	.3 Friendly sta	1 1719	8.3	51.4915712	-0.1963192					
Gran Via de les Corts Catalanes 668 Eixample 080	10 Ba Hotel Palace G	iL .			1 9	.4 The concier	1266	10	41.3916258	2.1716378					
la Fatebenefratelli 21 Milan City Center 20121 M	ilan I Hotel Cavour				1 8	.5 Shower plus	2463	7.5	45.4726942	9.1943911					
9 Vincent Square Westminster Borough London S	W1P Grange Roche	ster Hotel			0 8	.2 The room fu	1046	5.8	51.4942654	+0.1366026					
Shortlands Hammersmith and Fulham London W	6 8D Novotel Londo	n West			0 8	.3 Was really p	2443	8.3	51.4919588	-0.2200956					
02 220 Cromwell Road Kensington and Chelsea L	ondoi NH London Kei	nsington			0 8	2 No Negative	879	8.8	51.4948507	-0.1945544					
/ia Fatebenefratelli 21 Milan City Center 20121 M	ilan I Hotel Cavour	-			1 8	.5 Nothing	2463	9.6	45.4726942	9.1943911					
50 Oxford Street Westminster Borough London V	/1C 1 Radisson Blu B	dwardian Berkshire			0 8	.1 No Negative	1600	10	51.5146025	-0.1481978					
3 51 Wembley Hill Road Brent London HA9 8AU L					1 8	.9 Wifi didn t v	2274	8.3	51.5553678	-0.2864945					
More London Place Tooley Street Southwark Lon	don § Hilton London	Tower Bridge			0 8	.7 Cost a bit m	1705	8.3	51.504575	-0.082675					
i3 59 Kilburn High Road Maida Vale London Camd					0 7	.1 WINDOWLE	1877	3.8	51.5366053	-0.1919832					
7 Cromwell Road Kensington and Chelsea London	SW: Holiday Inn Lo	ndon Kensington Forun			0 7	.8 The hotel in	3867	7.1	51.4942305	-0.1851141					
IO trinity square City of London London EC3N 4DJ L	Jnite citizenM Towe	r of London			1 9	.1 Very comfy	4672	10	51.510237	-0.0764428					
eorge Street Westminster Borough London W1H	5BJ Durrants Hotel				0 8	.1 we were ple	e 1411	8.8	51.5179386	-0.1529794					
winguda Catedral 7 Ciutat Vella 08002 Barcelona	Spair Col n Hotel Ba	rcelona			1 8	.9 We had a co	1300	7.1	41.3849613	2.1756666					
alf Moon Street Westminster Borough London W	1J 7E Hilton London	Green Park			0 7	.3 No fridge in	1139	4.2	51.5064534	-0.1456146					
/iale Certosa 104 106 Certosa 20156 Milan Italy	Best Western					.5 The staff wa				9.1418864					
8 Albert Embankment Lambeth London SE1 7TJ U						.3 Large queue				-0.1214194					
6 avenue de Tourville 7th arr 75007 Paris France	Le Tourville Eit					.8 No Negative				2.3078116					
1 54 Buckingham Gate Westminster Borough Lor						.7 No Negative				-0.1376867					
57 Rue De Vaugirard 15th arr 75015 Paris France		Vaugirard Montparnas	e			.6 The toileter				2.3031011					
0 Lancaster Gate Westminster Borough London V						.7 Location	2400			-0.1795215					
2 Upper Woburn Place Camden London WC1H 0H						.9 No Negative				-0.1296571					
'ia Spadari 11 Milan City Center 20123 Milan Italy						.3 Perfect loca				9.1869092					
carsdale Place Kensington Kensington and Chelse			ton			.1 The staff w				+0.1917073					
puistraat 288 292 Amsterdam City Center 1012 V						.2 Great locati				4.8887815					
Vestminster Bridge Road Lambeth London SE1 7L			n			.7 Having stay			51.5009609						
26 Chiswick High St Chiswick London W4 5RY Uni						.5 No Negative				+0.2784701					
mpire Way Wembley Brent London HA9 8DS Unit						.3 Swimming			51.5590955						
Gracechurch Street City of London London EC3V a Rambia 128 Ciutat Vella 08002 Barcelona Spair						.2 Just the soft				-0.0847973					
a Kampia 128 Ciutat Vella 08002 Barcelona Spair	Hotel Serhs Ri	voii Kambla	_		U 8	.1 No Negative	1957	8.8	41.3843222	2.1712893			_	_	
WHOLE +															

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.

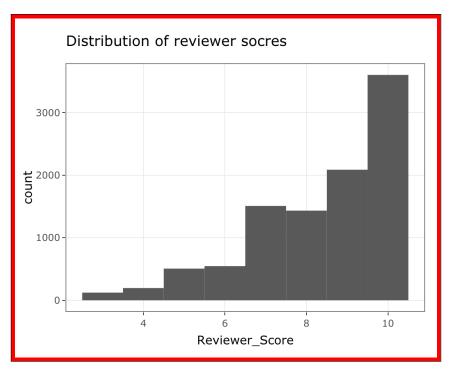


Figure 9: Reviewer_Score Plot

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

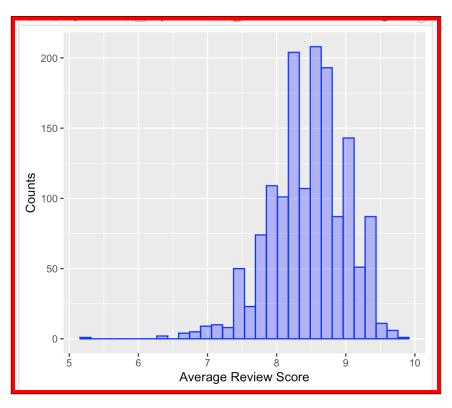


Figure 10: Average Review Score

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

score of the user ratings.

> data.fr	rame(p	percentile=perc,score=score)->d										
> print('	<pre>> print("Top rating scores are:")</pre>											
[1] "Top	[1] "Top rating scores are:"											
> d%>%arı	<pre>> d%>%arrange(desc(score))%>%head(5)</pre>											
percent	tile s	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

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

4 88400 0.088				antiplititi (0.0	。如何 的时候 0.08	ال الالالالا لا 0.0	airð iðiðiðið l 0.0	unistication of a 0.05	0.00	overp artic (0.0	арр ин й й й ц 0.06	^a Weinten I 0.0	min i (1997) (0.2	farli ni(iii) (0.08	de iliùijii 0.08		ans foiltinil 0.
(FR) (1996) 0.08	orp# ###### 0.06	199 0 (1997) (1 0.06	, Paraititin I, tr 0.9	ar@ #968899 4 0.06	rg elifetilitij (in 0.a	0.86	she rleysejjag (0.a	unwel8999991		te Attaining I	s () () () () 0.26	confi tional (0.08	re 461111	itseensiteiteet (0.00	899 400000 (a 0.06	ircon oniologi I 0.a	
196 799968 0.09	eñő fisiói (ac 0.00		"%%# 900000000000000000000000000000000000	upa ngisitin (0.08	400 00000 1 0.00	mią u (1990) (1 0.06	ox elew∰881 : 0.28	sound ia (1886) (0.28	1997 - 1 0.26	org anitik (1 a 0.0		.0.08	uncbirtition (refebi niski (u 0.00	nacq araithi (0.0	undagg inani i 0.2	B ritenini
10.00	e#9 96000000000000000000 00000000000000000	ा ग्रिसिसिंग 1.0	************** 6.00	re 69850000 L) (((())) 0.0	1 0.0	ciel i (1166) 0.06	ress ioner (0.0	1999990 U 0.40	sw ynnik (0.08	refresto ver i 0.06	Decision of the local sector of the local sect	ICONVERNMENT (0.08	(明朝朝)(0.3	.0.06	6 1444101001 0.0	
- confectifie	8 9994111 0.9	coffs idefield 0.96	···영화(1) (0.8	78 8866111111111111111111111111111111111	ap igibidi) (0.06	J Majaggoni BR.O	clety: 1000000000000000000000000000000000000		1.00	REALTING LINE	iograationi (0.a	, Markets en 0.0	reconstruction	neka kegi 1911 (ar 0.06	слик е байла Ц п. 0.06	Aylate ülüü (0.08	i Militerat, o
reco gnio (1966) (0.0	1.0.00	unayta shiri i 0.95	Sistematik I 0.86	coli ne (1 0.020	da#[]%()()()() 0.8	Id 報酬則 0.8	870005160 L	ace d an a hill I 0.09	해 해해당 0.06	cr \$190001 U 0.05	50 601118 0.8	runikkeiniinii I. 0.96	corts # MARKE I	resp 4000000000000000000000000000000000000	icasea skiik I 0.96		cər結婚期 C
657000000000000000000000000000000000000		excellent (1986 1010111111111111111111111111111111111	urestand in in	nciesp erities	s nag#111 (re 0.08	mb# <u>#99</u> 0000000000000000000000000000000000	ן אַנּוּנא נאיק 10.00 פּס.0	to osiyainii I 0.06	dis gi in the set of t	recti stiques (0.0	Jan Manageria	0.08		resap inini I 0.08	с8 (99)) () () 0.2	^{maja} jajang i c
^{screising (0.0}	im enti () (0.9	dev 85800000 1	micg@@@@##	n ulijenjije i re 0.0		exec ilitati i i h 0.8	ouseki 🏭 🗐 🛛 0.0	1.0.0	up ijijijiji j	sonstandill (0.0	a W akedining L 0.9	terni is/iii/iiiii l 0.2	sc Meilinini I 0.0	56973 61619 0.96	1 0.96	ь2 3%009) Ц а 0.96	
0.00	whategoogly I	19月1日 19月1日 19月1日 19月1日 19月1日 19月1日 1月月11日 1月111日 1月111日 1月1111 1月11111 1月1111111 11111111		aningini i	hag initisi i 0.028	1 100000000000000000000000000000000000	ir ssiliki i 68.0	e ji (11)()(1 1) 88.0	upo nisitiinii l 0.06		रक्षांग्रीम् । 0.0	жез ініція (0. 8	88# 100000000000000000000000000000000000	រក រិសមិសុទ្ធស្ន្រី re 0.06	corb anion II 0.a	Ridding I	-
a na) ((((()))))))))))))))))))))))))))))))	existinging i 0.0	1.000000000000000000000000000000000000	na%98)888	ež ina i		P r (明明) 月 0.0	199 6(1)(1) 190.0	c/8399994 4	19381101 L	uncon	ьэ н (жай) Ц 0.8	4090000 L	appr 351000 I	av sipējiļi ļ 0.0	fan tikkenikk i 0.06	95530 100 U 0.06	db illing i
\$\$\$\$1000 0.08	°90 01946101 I 0.089	ana ginaji (co 9 .06) ((())) 0.06	exp anilistic 0.0	ina şildiridi (0.08	uasilatika I o.a	78/1998 640100 4 0.00	fu spate(1000 0.08	na i Milijijiji 0.0	0.05	(1) (1) (1) (1) (1) (1) (1) (1) (1) (1) 	80-3 00/00 0.0	con8//////	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	transpaged I	4 /2006 1	in issisis
に 1993年1月 1993年11	다	br jalinijiji i	sundagija i	1.000 0.00	J HARSHER	аву ййж Ц 0.0		overch	916 39969 1 0.8	ch88 45014 0.08	a 65860000 I, 0.0	AREADIO I. 0.0		cle:#04666666666666 0.9	Desteoging (0.0	in trainint (re 0.0	
∿iensitäitää ∦⊂ 0.0) ()))))))))))))))))))))))))))))))))))	anf järjär i	neighbergigti I, 0.8	193日前前月 kr 0.8	1000 000000000000000000000000000000000	.6 %97400199 1 0.0	47.85 8888 8	be9秋期间 0.8	arc))))))))))))))))))))))))))))))))))))	imp ing the second second	pr 872904000 4 0.06	Jinferențiiții și o.a	영해 해하는 (0.0	"Exianalia 0.0	eceptioning i	。」) 第388894- 8.0	indianaliass o
coAsil080001 1, 0.8	at i (1996) (1996) (1 0.08	.0.06	personnel -	arc hianna I 0.06	1 (10) 91999999 0.0	con粉約期期 [0.0	1 1000000	ba胡晓树 (0.95		ung(2)00111 0.08	j (1999) 1.0	r神秘機構成 0.8	1.64499994 I	P isterini i 0.0	eX 188900 ا 0.0	9 0900000 1 0.08	arc hildiging 0.
atrastiticati I. 0.0	nso naisteineir L 0.0	transparentes (0.9	negative ; 0.8	iou stijaiaitti i 0.06	PHANDIAN 0.0	976/ 11111 (111	ا الاقترانية. 0. 8	्ञ /स्//स्/ । 80.08	Vielenniilii (0.a	REVIEW B	व ्यात्रात्रीय । re 0. a) Diddingan ew: 80.0	vi rening (0.06	sat en () 0.26	。 1) 動制的な の.9	40 120101101 (0.0	Net Managers
() (()()()() ()()()()()()()()()()()()()()('genii000 l 0.a	a #19.1911(1) a 0.06	ccentilitatii (r 0.09	edga ina (1 0.08	USERSING L	^u 8 ¢eAlebij8t I 0.06	م ا المُنْقَقَقَة 0.0) (Najan gero) 90.0	exc anin (0.8	aninamilia II 0.0	1.0 0	undergebietet I. 0.0	60100000000000000000000000000000000000	^{tr} ite s (1) 0.0	inggindin I 0.0	manqau xila i s 0.0	sup rovina
rias cineteri I, 0.06		cons adilitita (0.0	0.08						tf_idf								

Figure 13: Topic Modelling

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



Figure 14: Countries mentioned in dataset

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



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.

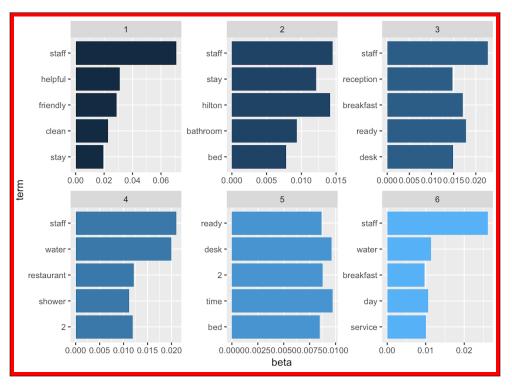


Figure 16: TF-IDF

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

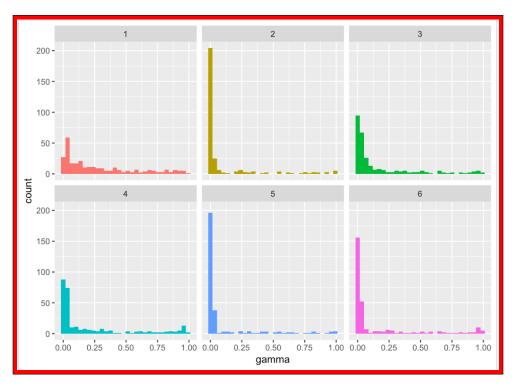


Figure 17: Topic Modelling

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 algorithm.

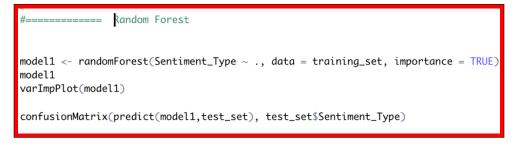




Figure 20 shows the ANOVA results using ChiSquare Test.

> 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 *** <2e-16 *** Total_Number_of_Reviews 1 363.4 6997 5703.6 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 <- predict(on(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</pre>
```

