IS5152 – DECISION MAKING TECHNOLOGIES

Final Project

Group 8:

CHEN Changhua	**A0039923Y**
Lu HE	**A0160883Y**
YANG Kan	**A0161013B**
YAU Chung Yin	**A0160936B**

Lecturer: **Prof HUANG Ke Wei**

Master of Science in Business Analytics (MSBA)

National University of Singapore

AY2016/2017 Semester 2

Contents

Abstract

This report aims to investigate the major drivers affecting customers' review ratings on Starbucks (and on Dunkin' Donuts as a comparison) based on Yelp Dataset Challenge 2017, with focus on textual analysis on the review and tip generated by customers. **Panel linear model with location clustering as index** is used as the main model with comparison from ordinary least squares (OLS), Rpart, Random Forest (RF) and Generalized Boosted Regression Models (GBM). The main model gives R^2=0.662. Variables are derived from methodologies including k-means clustering, dictionary approach, topic modelling (by Latent Dirichlet Allocation), text classification (by SVM and Random Forest). Main model has been further enhanced via various imputation methods of missing values. The insights revealed in this report lead to several recommendations which not only be useful for Starbuck's management, but also other food chain affiliations.

Key Words: PLM, RF, OLS, Rpart, GBM, SVM, LDA, K-mean clustering, Word Cloud, Dictionary approach, topic modeling, text classification

1. Overview

Starbucks, with a total number of company-operated and licensed stores of 25,085 over the globe as of 2 October 2016 (Starbucks Corporation 2016a), is a leading player in the coffee and tea segment of the food & beverage industry. Despite the decent growth in revenue from 2015 to 2016 by 11% which was partially contributed by the increased numbers of stores in operations, the growth rate in comparable store sales - a key performance indicator for retailers' health - is decreasing slowly, from 7% in 2015 to 5% in 2016, with latest growth in 3% in the first quarter in 2017 (Starbucks Corporation 2016b).

Meanwhile, Dunkin' Donuts, one of Starbucks' competitors especially in the United States, has aggressively tapped into the coffee war by product innovation, leading beyond Starbucks in terms of share price performance in the past year (Monica 2017).

From the long-term perspective, Starbucks' management need to assess carefully about how to maintain customer loyalty when the organic growth in market becomes stagnant and competition intensifies. "Customers choose among specialty coffee retailers primarily on the basis of product quality, service and convenience, as well as price" (Starbucks Corporation 2016a), and customers' feedback from external sources, like Yelp, becomes vital for Starbuck's management to understand what keep customers with them and what drive them away.

Since its establishment in 2004, Yelp has expanded rapidly and become one of the major online business review platforms in the United States. Previous studies provide evidence of the positive correlation between restaurants' revenues and consumers' review rating on Yelp. The benefits are obvious on independent restaurants but not on restaurants with chain affiliation. In fact, the market shares of chain restaurants declined as Yelp penetration increased (Luca 2011). Another study on the credibility of Yelp reviews reveals that chain restaurants are less likely to commit review frauds, i.e. leaving positive fake reviews (Luca and Zervas 2013). For Starbucks, the existence of Yelp reviews has a two-sided impact: Yelp makes information about independent coffee shops more easily accessible to customers and benefit their revenue streams as mentioned above; it also provides a pool of relatively non-biased customers' feedback to better understand the market. The actual data used in this study also shows significantly lower overall review ratings on chain stores (both Starbucks and Dunkin' Donuts) as detailed in Section 2.1.

2. Data

This study is based on the most recent 2-year Yelp reviews available from 20 January 2015 to 20 January 2017. It is assumed that in food industries, only recent reviews have high relevancy to customers and the restaurant management. Among the total 1,962,391 reviews (referred to as "all" or "the universe") on 121,918 restaurants or other service providers, 7,739 reviews pertain to 806 Starbucks stores over 110 cities and constitute the main subject of interests. Comparatively, there are 1,936 reviews Dunkin' Donuts which covers 165 stores in 66 cities during the same period.

2.1 Dependent Variable

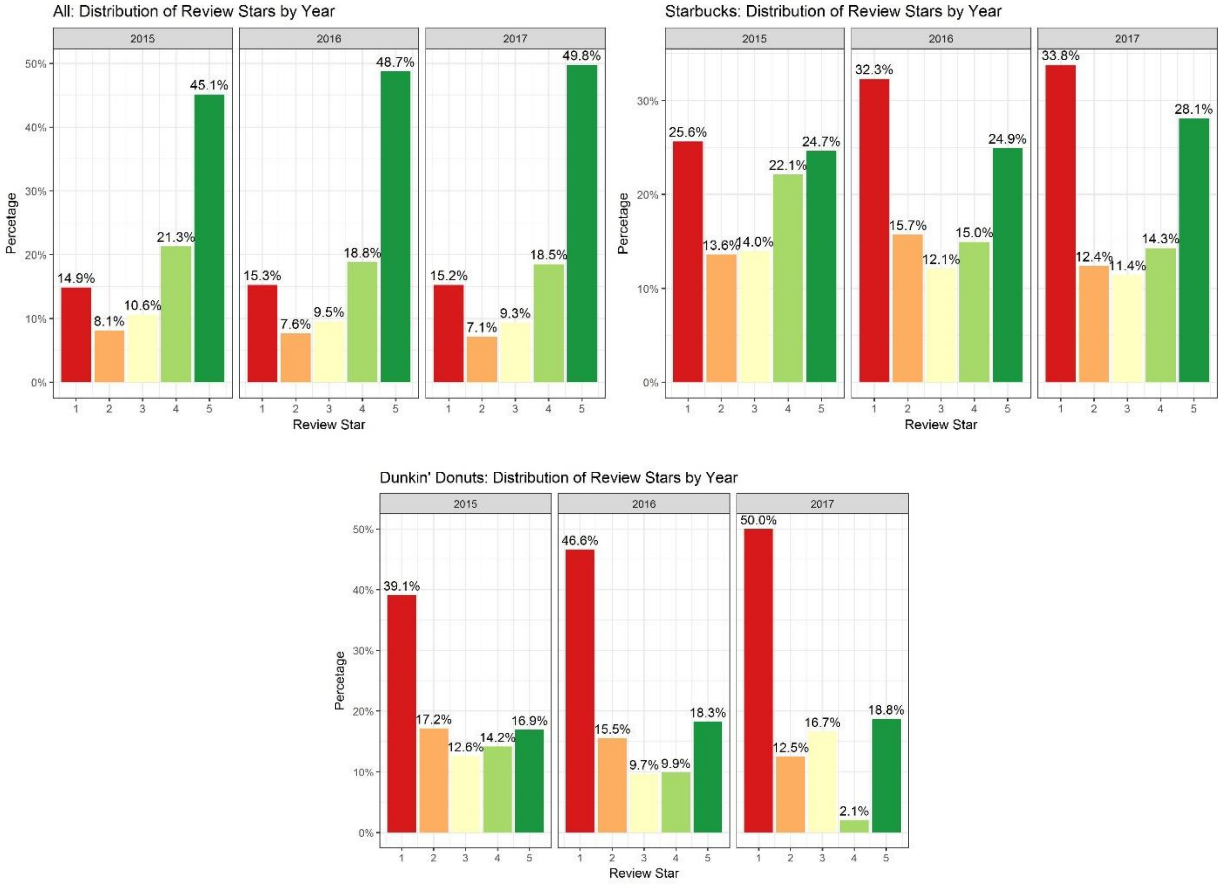
The dependent variable of interest is the review rating (star) (on a discrete scale from 1 to 5) given by the Yelp users and the summary statistics are shown in **Table 1** below:

Table 1: Summary Statistics of Review Rating

	Number of Reviews	1st Quantile	Median	Mean	3rd Quantile	Standard Deviation
All	1,962,391	3.00	4.00	3.76	5.00	1.48
Starbucks	7,739	3.00	3.00	2.95	4.00	1.58
Dunkin' Donuts	1,936	1.00	2.00	2.44	4.00	1.55

Review ratings on Starbucks are generally lower on average as compared to the universe as evidenced by a lower star in median, mean and 3rd quantile. **Figure 1** shows that the 5-star reviews constitute the largest group in the universe; while for Starbucks, 1-star review not only constituted the largest rating class, but also rose rapidly from 25.6% in 2015 to 32.3% in 2016. Donkin's Donut also display a similar distribution with even larger percentage of 1-star reviews.

Figure 1: Distribution of Review Stars by Year (All, Starbucks and Dunkin' Donuts)



For Starbucks, reviews are highly concentrated in Nevada and Arizona (the United States) and Ontario (Canada), mainly contributed by Las Vegas (28.05%, 125 stores), Phoenix (14.05%, 71 stores) and Toronto (5.70%, 116 stores). In addition, stores in Las Vegas and Phoenix are the main contributors of lower-starred review ratings as shown in **Figure 2** and **Table 2** below:

Figure 2: Starbucks - Distribution of Review Stars by State and City

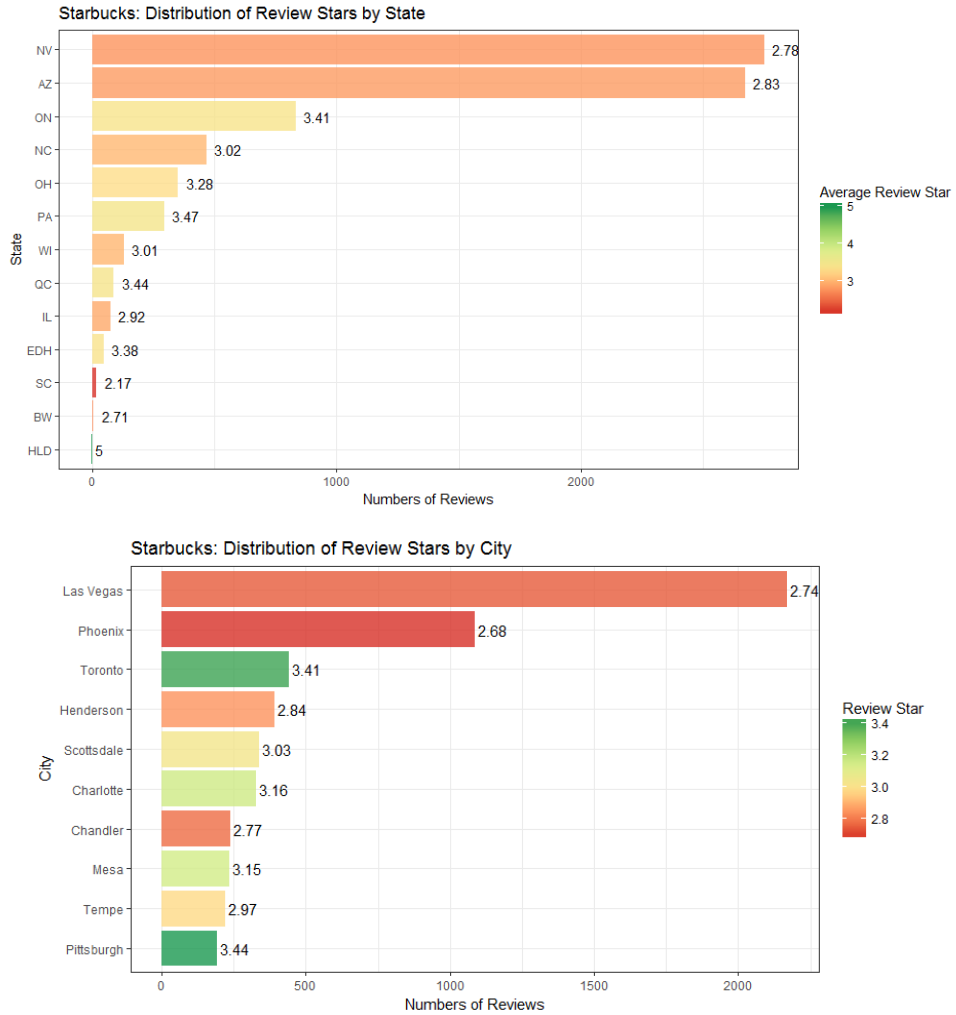
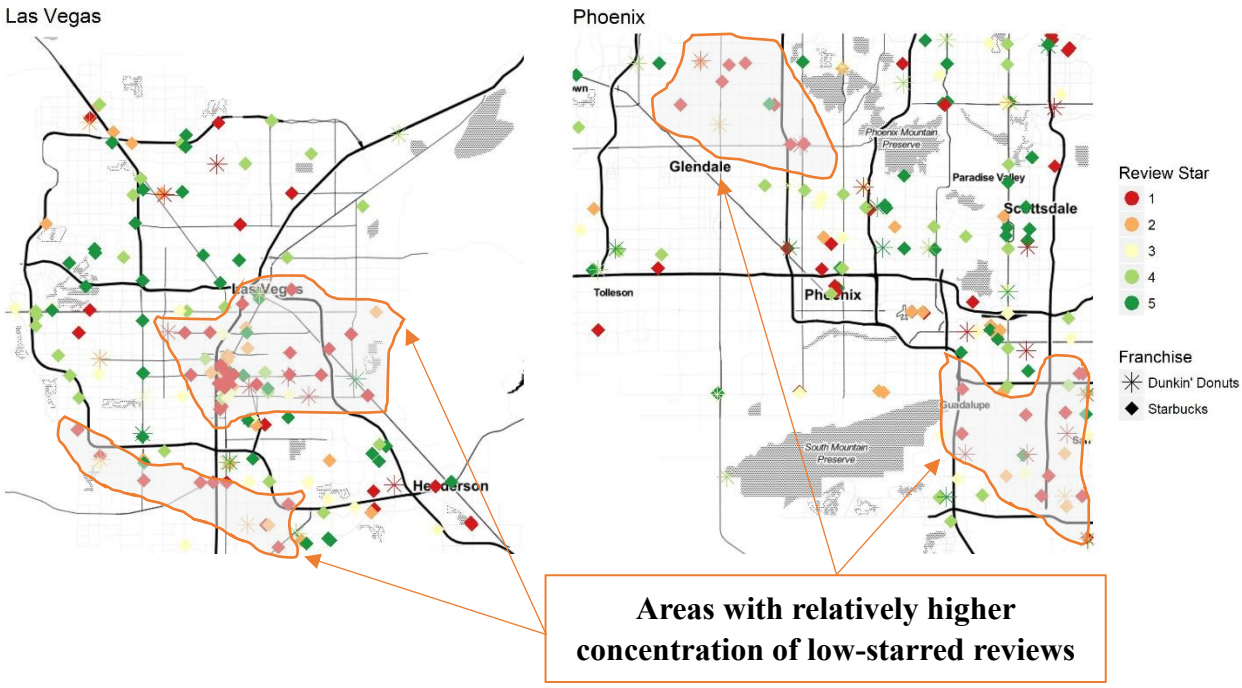


Table 2: Starbucks - Summary Statistics of Review Rating (Las Vegas and Phoenix vs others)

	Number of Reviews	1st Quantile	Median	Mean	3rd Quantile	Standard Deviation
Las Vegas & Phoenix	3,258	1.00	2.00	2.72	4.00	1.60
Others	4,481	2.00	3.00	3.11	5.00	1.54

Figure 3 further visualizes geographically the reviews on Starbucks (with Dunkin' Donuts as comparison) in Las Vegas and Phoenix. Certain areas seem to have a higher concentration of lower-starred reviews:

Figure 3: Geographic Distribution of Review Stars in Las Vegas & Phoenix



Section 2.2.2 further explored the potential impact of location on review ratings.

2.2 Key Features

2.2.1 Textual Informaiton

User-generated textual information from *Review* and *Tip* written by Yelp users are the main feature of interest in this study. *Review Text* has the same granularity as review rating (star) and therefore can be regarded as a qualitative description that the sentiment of which directly correlates with the review rating given by users. *Tip Text*, which also provides customers' sentiment on their dining experience, does not come with a quantitative point of scale.

Exploratory (word cloud) analysis is performed among two types of textual information: a) 5-star reviews vs 1-star reviews on Starbucks; and b) tips on Starbucks stores on high-performing stores (average review ratings of 4 or above) vs low-performing stores (average review ratings of 2 or below), which provides an overview of key words frequently-mentioned by Yelp users.

Three ways of counting frequency of words are used in generating word clouds: unigram, bigram and trigram. On *Review Text,* **trigram** gives the most intuitive and detailed views among the three and is adopted. On *Tip Text,* **bigram** gives a better illustration as the output from trigram is too sparse. For further details, please refer to Section 5.2.2 for the discussion on enhancement/improvement on generating word clouds.

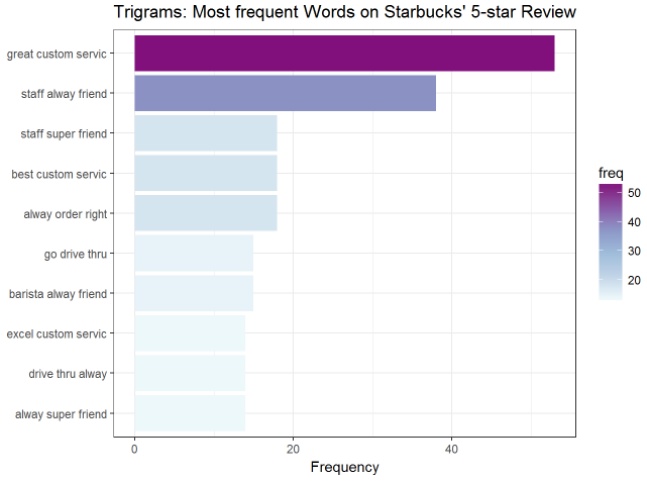
Figure 4 shows that and "customer service" are the common top frequently-mentioned topic in both 5-star (as "great customer service" and "best customer service") and 1-star reviews (as "worst customer service", "worst service ever", "horrible customer service", "poor customer service").

Among 5-star reviews, friendliness of staff is a noticeable topic (indicated by "staff always friend", "staff super friend", "barista always friend", "always super friend", etc.). Order execution (indicated by "always order right") is also a high ranked topic among 5-star reviews.

Among 1-star reviews, frequently-mentioned phrases include "slowest drive through", "wait drive through" and negative statements such as "never go back".

Figure 4: Word Cloud on Reviews on Starbucks

5-Star Reviews



1-Star Reviews

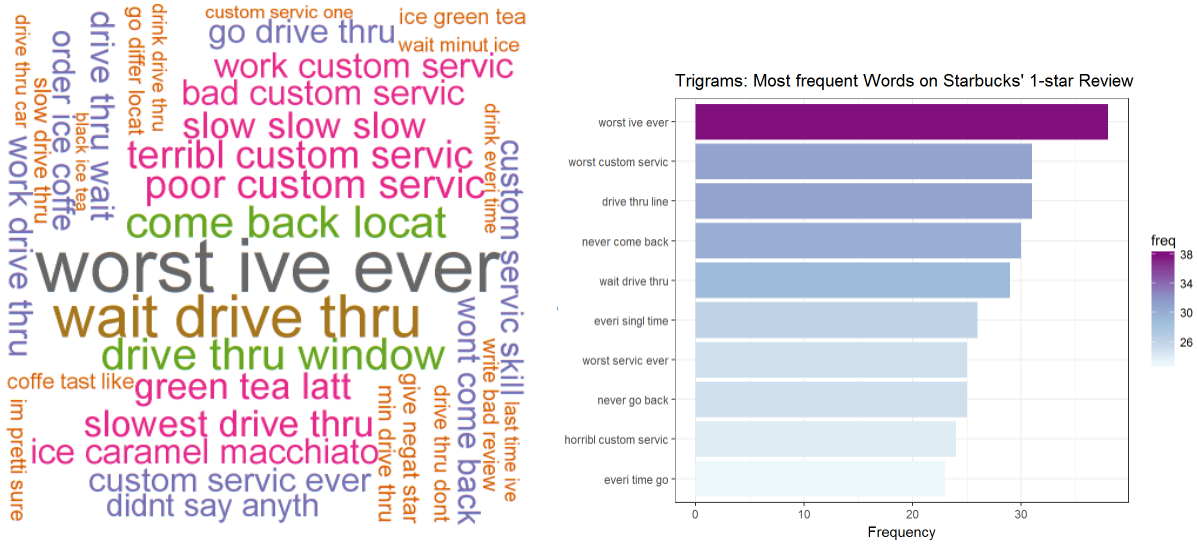
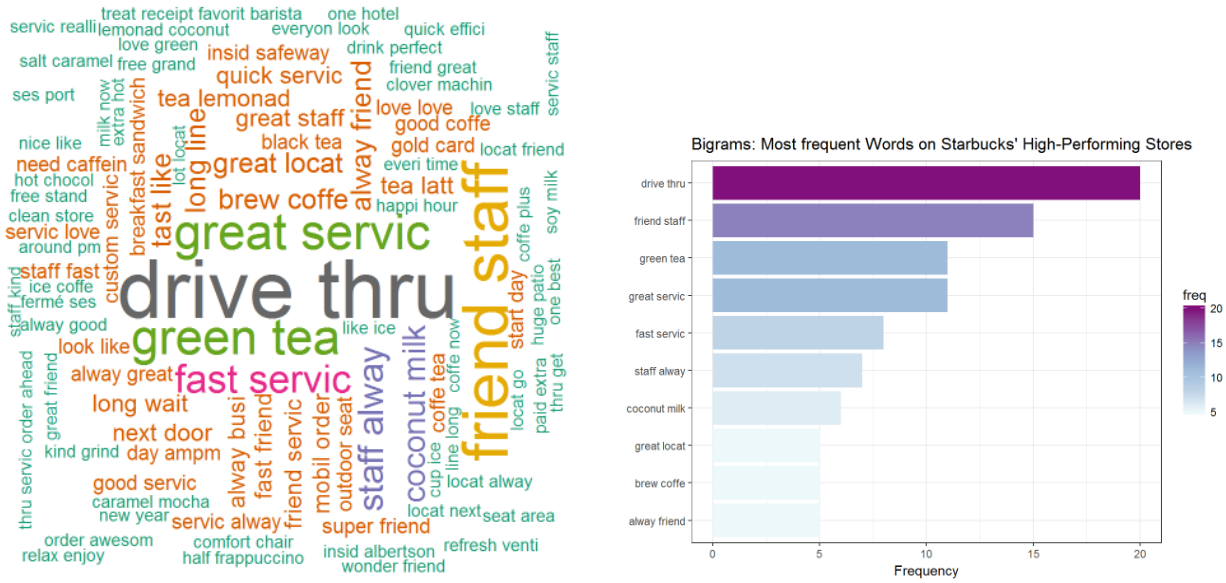


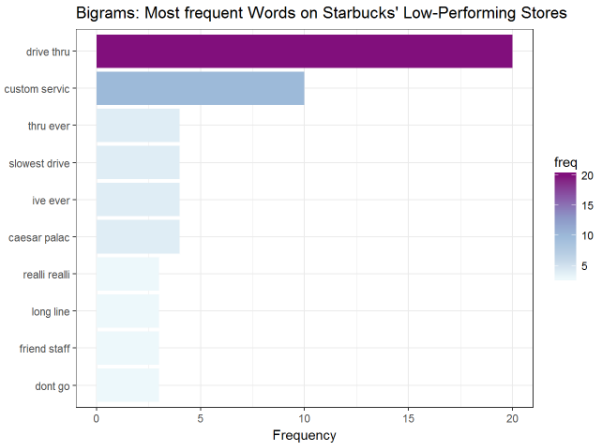
Figure 5 further gives hints some "recommended" items on high-performing stores, such as "green tea" and "coconut milk"; and the "unsatisfactory" items about low-performing stores, such as "slowest drive".

Figure 5: Word Cloud on Tips on Starbucks

Tips on High-Performing Stores



Tips on Low-Performing Stores



Word clouds provide an intuitive view about what is being frequently mentioned, while quantitative and analytical approaches are necessary to construct variables that extract most meaningful textual information to understand the customers' view as detailed in **Table 3**.

Table 3: Variables Created from Textual Information

Variables Name	Type	Methodology	Section
Sentiment scores	Numeric	Dictionary Approach	3.2.1
Binary labels	Categorical	Text Classification (SVM)	3.2.2
Topics and probability	Categorical/Numeric	Latent Dirichlet Allocation (LDA)	3.2.3

These three types of variables are moderated to present the intrinsic patterns and characteristic of text mining, and form independent variables in main model.

Please refer to the specific sections on the methodology of how each type of variable is derived.

2.2.2 Restaurant-Level Variables

Figure 3 provides intuition that location/neighbourhood could be a potential restaurant-level variable that have an impact on review rating. There are 2 variables in the existing variables in Yelp data that can be used as categorical proxy of location: *Postal Code* and *Neighbourhood*:

Table 4: Summary of Existing Location Proxy

Franchise	Numbers of Stores	Variable Name	Levels	Missing Values
Starbucks	806	Postal Code	487	3
		Neighborhood	121	499
Dunkin' Donuts	165	Postal Code	134	1
		Neighborhood	27	128

Table 4 above shows that *Postal Code* has minimal number of missing values, the number of factor levels, however, is relatively high as compared to the numbers of stores (less than 2 stores per postal code on average) for both franchise. For *Neighbourhood,* more than 50% values are missing from the dataset.

Hence, a location clustering is done based on the *Longitude and Latitude* provided in the dataset cluster in Las Vegas and Phoenix, cities identified as major contributors of lower-starred reviews. Please refer to Section 3.2.4 for the methodology.

2.2.3 User-Level Variables

Users' individual attributes are also relevant factors that might affect review rating, these include:

1. Personal strictness, indicated by average stars given by the individua user;
2. Familiarity/heaviness of usage with Yelp, indicated by the number of fans, number of total review made, etc.; and
3. Perceptions from other users, such as compliments received.
4. Relative active level of user compared with other users. Active level is determined by all compliment related features in user profile such as useful, compliment photos, compliment list, compliment funny, compliment plain, compliment note, compliment count, compliment writer, compliment cute etc.

All such user-specific variables are also considered together with the textual information on reviews.

3. Methodology

3.1 Main Model

With all the dependent variables introduced in Table 3, the key business question for this study is to investigate what factors affect the review rating on Starbucks to provide insights on Starbucks' major area of improvement in the future. The final data used for the main model also include observations on Dunkin' Donuts to provide a broader view from customers on Starbucks and one of its competitor as a whole.

Rating ~ numerical and categorical variables (original datasets)
+ numerical variable of sentimental scores (Section 3.2.1)
+ categorical topics variables annotated from text classification (service, food, ambience, facilities available and values add to customer, Section 3.2.2)
+ categorical variable and numeric probabilities of topics derived from topic modeling (Section 3.2.3)

The main model is by Panel Linear Model (PLM with business location as index) while other methodology was also tested: Ordinary Least Square (OLS) linear regression is initially used as the main evaluator to investigate the significance of each attributes towards review rating; Decision tree (Rpart) and Random Forest (RF) are used to find out the main contributors to the rating to check against the results from OLS and PLM. Based line results are evaluated in detail in Section 4.

Handling Missing Value

Among of the variables created, *Sentiment Score 2 ("Score2")* and its derived variables, *Score2_very positive* and *Score2_very negative*, have 179 records of missing values. Since *Score2 (*together with *Score2_very positive* and *Score2_very negative)* are one of the important attributes explaining review rating (see Section 4), 5 methods of imputation were performed**.** Based on the evaluation across the 5 methods, **Method 3 which impute *Score2* by its monthly average across all observations** resulted in the best performance and was adopted. For the evaluation details,

please refer to Section 5.1.1. With the imputed *Score2*, *Score2_very positive* and *Score2_very negative* are binary variables using the 3rd and the 1st quantile as cutoff as further explained in Section 3.2.1.

3.2 New Variables

3.2.1 Sentiment Score (Dictionary Approach)

Dictionary approach was adopted in creating a *sentiment score* to provide a quantitative indicator on the review text based on the number of counts of "positive" and "negative" words referring to the pre-defined dictionaries, which contained positive and negative words of 2,006 and 4,783, respectively.

Among the three variations in calculating sentiment scores, **Formula 2** was adopted, in which on each individual review, the sentiment score was calculated as the percentage of numbers of matched-positive over the total numbers of matched-positive and matched-negative, so that the sentiment score will be on scale from 0 - 1:

$$score2_i = \frac{\text{positive.matches}_i}{\text{positive.matches}_i + \text{negative.matches}_i} \quad \text{where } i \in \{review\_id\}$$

Scores exceeding the 3rd quantile are considered as "very positive", and those below the 1st quantile are considered as "very negative", created as binary variables.

Please refer to Section 5.2.1 for the details on **Formula 1** and **Formula 3** and the evaluations among the 3 formulas.

Table 5: Summary of Results: Sentiment Scores on Review Text

	Formula 1		Formula 2		Formula 3	
	Starbucks	**Dunkin' Donuts**	**Starbucks**	**Dunkin' Donuts**	**Starbucks**	**Dunkin' Donuts**
Min	-14.00	-16.00	0.00	0.00	-8.23	-7.75
1ST Quantile	-1.00	-1.00	0.43	0.36	-0.72	-0.70
Median	2.00	1.00	0.67	0.61	0.06	0.00
Mean	1.82	1.29	0.64	0.59	0.00	0.00
3rd Quantile	4.00	3.00	1.00	0.86	0.77	0.81
Max	24.00	22.00	1.00	1.00	4.97	5.92
Missing Values	-	-	155	24	-	-
"Very Positive"	2,186	643	2,060	497	1,915	422
"Very Negative"	1,233	391	1,811	478	1,747	474

3.2.2 Text Classification

Text classification is supervised learning by using a training dataset (with manually annotated labels) to train a model to predict other observations. 200 records are selected out of 9,675 reviews for manual labelling, with 150 from Starbucks and 50 from Dunkin' Donuts. The reviews were read manually with human interpretation of the meaning and label them in the following aspects:

1. Service
    a. If customer is getting the correct order he/she made.
    b. If the queue time and serving time is within a reasonable range, and acceptable by customer.
    c. If the staff are friendly and customer felt welcome
    d. If the location always has stock of what customer ordered.
    e. If customers are charged with the correct amount.
2. Facility
    a. If the location is easily accessible, such as convenient parking, able to drive thru, or located in good places.
    b. If the location has WIFI, and it works well
    c. If the location supports various payment method, including App, mobile etc.
    d. If customer is always able to find a seat
    e. If the toilet is tidy and clean.
3. Food
    a. If the food is tasty
    b. If the food has good quality, i.e. meets the standards as it is expected to be
    c. If the location has a good variety of selections
    d. If the food the fresh
4. Ambience
    a. If customer has enjoyed being there.
    b. If the tables, chairs, floors are clean.
    c. If the decoration is cozy, and seat arrangement is good.
    d. If the environment is quiet, or probably with good music.
5. Value

a. If the product and service is worth the money customer is paying.
b. If customer can get a good discount with their loyalty program.
c. If customer sometimes get freebies

For manual labelling, efforts are focused to extract as much as information from the reviews to provide a good training data set to the model. 3 labels are used per each variable and review, they are positive, negative, or not mentioned. If it is positive, 1 will be labelled, if negative, -1 will be labelled, otherwise 0. These labels may not be directly used in the text classification modelling, as we could also create new labels using the preliminary information. It will be explained later.

Since our final model is to use "Star" is dependent variable, and analyze the reason of why customer rate the business high or low. These variables are constructed for such a purpose. If any these variable is more significant than the others in the final model, which will mean the customer cares about such variable more than the rest. The business could try to improve this aspect to improve their ratings in Yelp. Eventually, the business could use this information to attract more new customers and retain the existing ones.

As mentioned above, there are a few different ways of constructing the new variables. It is listed below:

1. **3 Labels (Positive, negative, not mentioned):** This is to use the manually labels directly. This requires to train the model as a multi-label classification. This would be the best for the final model as it gives the most information of the text. However, with just 200 observation training set, the prediction for the rest of the test observations may not be good.
2. **2 labels (Positive, Not positive):** This is to merge "negative" and "not mentioned" into the same label. This label will tell if customers have a good feedback on a location or not. Since this is a binary classification, the performance may be better than the previous way.
3. **2 labels (Negative, Not negative):** This is to merge "Positive" and "not mentioned" into the same label. This label will tell if customers have a bad feedback on a location or not. This is again a binary classification. Model selection will be based on predict performance.

Random Forest performed the best as compared to SVM. Please refer to Section 5.2.2 for the evaluation of results.

Those 200 training observations and the rest of the observations are grouped into the same data frame. The command "train_model" from library "RTextTools" is used to train the model by random forest. The command "classify_model" from the same library is used to predict the dependent variables. There are total 5 model build, and they are used for predicting the label "Bad Service", "Bad Food", "Bad Ambience", "Bad Facility" and "Bad Value" on each of the review. These predicted labels are added back to the original dataset. The primary key of this dataset is at "Review_ID". This will also be used to merge with the other variables for the final model.

The performance of the text classification can be potentially enhanced by manually labelling more reviews as the training set. It will provide the model more training data and the predicted results will be more accurate. If the number of labelled reviews is large enough, doing a 3-label classification will provide the final model a set of more meaningful variables.

3.2.3 Topic Modelling

LDA models are tuned with number of topic (k), and AIC is compared to decide whether the assigned topics or the associated probabilities with each topic assignment would be adopted as the new variables for the main model. After evaluation, the associated probabilities with each topic assignment from the LDA model with parameter k = 10 generated the lower AIC, therefore they would be added as the new variables for the main model.

Please refer to Section 5.2.3 for the details on evaluations of LDA Models.

Figure 6: 10 Topics from LDA Model

```
##       Topic 1 Topic 2    Topic 3  Topic 4    Topic 5   Topic 6 Topic 7
##  [1,] "coffe" "place"    "locat"  "alway"    "servic"  "drive" "back"
##  [2,] "ice"   "seat"     "work"   "friend"   "custom"  "wait"  "ask"
##  [3,] "cup"   "lot"      "morn"   "staff"    "employe" "line"  "walk"
##  [4,] "tea"   "park"     "usual"  "love"     "manag"   "minut" "told"
##  [5,] "latt"  "nice"     "realli" "starbuck" "bad"     "thru"  "counter"
##  [6,] "tast"  "area"     "close"  "locat"    "rude"    "peopl" "girl"
##  [7,] "cream" "starbuck" "time"   "quick"    "worst"   "long"  "guy"
##  [8,] "hot"   "tabl"     "come"   "best"     "experi"  "time"  "came"
##  [9,] "milk"  "locat"    "busi"   "fast"     "store"   "slow"  "look"
## [10,] "extra" "sit"      "get"    "nice"     "place"   "insid" "hand"
##       Topic 8    Topic 9   Topic 10
##  [1,] "starbuck" "order"   "coffe"
##  [2,] "locat"    "drink"   "donut"
##  [3,] "visit"    "time"    "dunkin"
##  [4,] "free"     "barista" "food"
##  [5,] "star"     "wrong"   "tri"
##  [6,] "price"    "everi"   "sandwich"
##  [7,] "card"     "make"    "place"
##  [8,] "store"    "correct" "better"
##  [9,] "expect"   "mess"    "back"
## [10,] "next"     "els"     "breakfast"
```

From the above figure, topics of the reviews are diverse. Topic 1 and 10 are related to food and drinks reviews, and reviews under Topic 4, 6 and 9 are about efficiency. Other topics include location for topic 2, service for topic 5, and people and staff for topic 4 and 7. Various elements are included in Topic 3 and 8 such as price and location.

Table 6: Summary of the 10 topics

Categories	Topics
Food and Drinks	1, 10
Efficiency (Positive)	4
Efficiency (Negative)	6, 9
Service (Negative)	5
Location	2
Staff (Positive)	4
Staff (Negative)	7
Other	3, 8

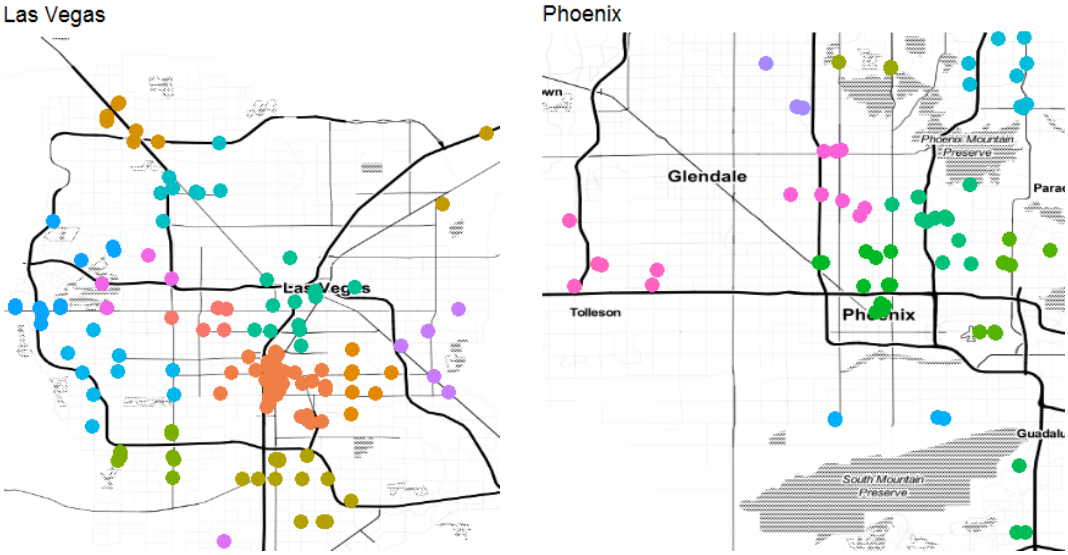
3.2.4 Location Clustering

Location clustering was performed by the following steps:

1) Calculate the pairwise distances between all points given (based on the given *Longitude* and *Latitude* in the Yelp dataset) and return either a distance or full matrix as specified;
2) Get the shortest distance between two points according to the "haversine method";
3) Use k-means clustering based on the computed distance by specifying k = 28 on all data points relating to Las Vegas and Phoenix (as discussed in Section 2.2.2); and
4) All other cities are labeled as a separate category (location 0).

The value of k was arbitrarily chosen to reach an output that can cluster the stores in Las Vegas and Phoenix with reasonable level of details. A range from 20 to 30 makes intuitive sense and the output based on k = 28 is shown in **Figure 7** below:

Figure 7: Location Clustering Result on Las Vegas and Phoenix



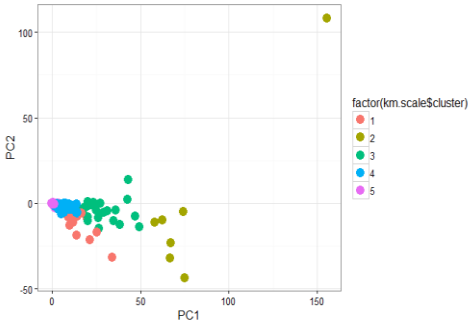
Including "location 0" that is labeled to all other cities, the "Location Clustering" label of 29 categories are created.

3.2.5 User Clustering

User clustering is performed using PCA and K-means to cluster the user profiles in unsupervised model.

Firstly, user profile dataset has been prepared from user dataset, keep those quantity variables and substrate other variables such as user id, name, date and elite attribute. Secondly, scale those quantity variables in range [0,1] by dividing each value by maximum value in the same column. Thirdly, using k-means to cluster the users into 5 clusters, corresponding to 5 star ratings. Result is shown in **Figure 8** below. Thus, k-means clustering on user profile serves as a newly created variable for possible model improvement later.

Figure 8: K-mean clustering and PCA plot




```
table(user_profile$kmscale.cluster)

   1    2    3    4    5
  14    7 7710  203   24
```

4. Baseline Results

The main model is generated over various stages and with various combination of variables. We have applied various technics to understand which factors statistically influence the star rating, ranging from OLS, PLM, Rpart, Random Forest (RF) and GBM. The baseline results and interpretation from OLS and PLM is shown below and please refer to Section 5.1.2 for Rpart, RF and GBM.

Stage 1: Using OLS method, regress stars rating on created variables obtained in 3 types of text analysis: Sentimental scores, extreme positive and negative score based on 3 variations; 10 topics based on topic modeling; some attributes from user dataset to find any individual preferences; and 28 location clustering for business of Starbuck, as well as 5 main indicators obtained from text classification.

Result tabulated in **Table 7** column (1) shows that sentimental scores in general are statistically significant. However, the sentimental scores with formula 2 provides more consistent results on overall sentimental score, extreme positive and negative. Topic 4, 5, 6, 8, 9 and 10 are more statistically significant than other topics. Number of fans the user have also contribute significantly to the particular rating. In term of binary relevance, service and food are significant factors contributing to star rating. This is not surprising.

The baseline of OLS has provided $R^2 = 0.542$

Stage 2: Using StepAIC to select the best model and compare with the result obtained in Stage 1. Result tabulated in **Table 7** column (2) shows the $R^2 = 0.542$ also and the significant variables are quite consistent with those identified in Stage 1.

Stage 3: Vary the model with different scores formula, which can be mutually exclusive, as well as topic and topic probability. Results in **Table 7** column 3 to 8 shows that the outcome of model variation using different combination of score formulas and topics. **By consider only one sentiment score formula and one set of variable from LDA only, column 4 (score2 and using LDA probabilities) provide the best results,** with $R^2 = 0.522$.

Stage 4: To further investigate the effect of location clustering on review, we use panel linear model (PLM) on the main model and variations. Results in **Table 8** shows that in general, R^2 of each corresponding model increases. The best model **(consider using sentiment score created by one formula only)** using PLM is column (4) $R^2 = 0.560$. This shows that sentimental score using formula 2 provide more significant influence on star rating.

Stage 5: Imputation of missing values on the best model stated in Stage 4 can further enhance the performance. Various imputation methods have been explored (Section 5.1.1). Replace missing value with month average lead to better performance and enhance the best model with **$R^2 = 0.662$**, further increase from 0.560 previously. Refer to **Table 9** for details.

Table 7: Comparison of OLS results

OLS - Results Comparison

	Review Stars							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
score1	0.133***	0.125***	0.151***			0.161***		
	(0.012)	(0.012)	(0.007)			(0.007)		
score1_very.pos1	0.087	0.099*	0.268***			0.282***		
	(0.056)	(0.056)	(0.042)			(0.042)		
score1_very.neg1	0.087		-0.249***			-0.235***		
	(0.062)		(0.043)			(0.043)		
score2	1.912***	1.940***		2.832***			2.960***	
	(0.107)	(0.106)		(0.095)			(0.095)	
score2_very.pos1	0.186***	0.174***		-0.041			-0.069	
	(0.046)	(0.045)		(0.043)			(0.043)	
score2_very.neg1	0.133**	0.148***		0.244***			0.254***	
	(0.059)	(0.053)		(0.052)			(0.053)	
score3	-0.245***	-0.218***			0.388***			0.427***
	(0.036)	(0.031)			(0.020)			(0.020)
score3_very.pos1	0.104*	0.095*			0.485***			0.503***
	(0.058)	(0.057)			(0.041)			(0.041)
score3_very.neg1	-0.100*				-0.195***			-0.158***
	(0.056)				(0.045)			(0.045)
R_LDA_topic2	-0.029	-0.028				-0.095*	-0.113**	-0.123**
	(0.067)	(0.067)				(0.056)	(0.055)	(0.057)
R_LDA_topic3	0.071	0.072				0.573***	0.487***	0.523***
	(0.062)	(0.062)				(0.051)	(0.050)	(0.052)
R_LDA_topic4	-0.285***	-0.307***				-0.301***	-0.265***	-0.300***
	(0.069)	(0.064)				(0.057)	(0.056)	(0.058)
R_LDA_topic5	-0.308***	-0.308***				-0.401***	-0.364***	-0.412***
	(0.068)	(0.068)				(0.057)	(0.055)	(0.057)
R_LDA_topic6	-0.191***	-0.189***				-0.453***	-0.413***	-0.446***
	(0.068)	(0.068)				(0.053)	(0.052)	(0.054)
R_LDA_topic7	-0.018	-0.015				0.100*	0.151***	0.139***
	(0.068)	(0.061)				(0.053)	(0.052)	(0.054)
R_LDA_topic8	-0.194***	-0.194***				-0.372***	-0.335***	-0.378***
	(0.071)	(0.071)				(0.060)	(0.059)	(0.061)
R_LDA_topic9	-0.251***	-0.254***				-0.508***	-0.427***	-0.466***
	(0.072)	(0.072)				(0.059)	(0.057)	(0.059)
R_LDA_topic10	-0.274***	-0.249***				-0.268***	-0.189***	-0.343***
	(0.068)	(0.061)				(0.055)	(0.053)	(0.055)
R_Topic1_prob	-1.296**	-0.983*	0.326	0.741	1.422***			
	(0.623)	(0.509)	(0.497)	(0.486)	(0.503)			
R_Topic2_prob	-3.006***	-2.713***	-1.439***	-1.434***	-0.708			
	(0.614)	(0.526)	(0.489)	(0.471)	(0.494)			
R_Topic3_prob	4.935***	5.253***	8.968***	8.346***	9.297***			
	(0.623)	(0.517)	(0.473)	(0.451)	(0.478)			
R_Topic4_prob	-0.620		-0.258	-0.117	0.616			
	(0.599)		(0.468)	(0.450)	(0.473)			
R_Topic5_prob	-1.758***	-1.411***	-1.308***	-1.525***	-0.477			
	(0.610)	(0.491)	(0.473)	(0.453)	(0.481)			
R_Topic6_prob	-3.020***	-2.743***	-2.248***	-2.136***	-1.463***			
	(0.540)	(0.441)	(0.394)	(0.380)	(0.399)			
R_Topic7_prob	-0.306		1.373***	2.024***	2.645***			
	(0.584)		(0.424)	(0.407)	(0.427)			
R_Topic8_prob	-2.196***	-1.855***	-1.154**	-1.276***	-0.224			
	(0.629)	(0.481)	(0.497)	(0.477)	(0.504)			
R_Topic9_prob	-2.932***	-2.583***	-1.928***	-1.813***	-0.600			
	(0.606)	(0.480)	(0.455)	(0.437)	(0.460)			
R_Topic10_prob								
U_useful	-0.00001		-0.00002*	-0.00000	-0.00002	-0.00002*	-0.00000	-0.00002
	(0.00001)		(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
U_fans	0.001***	0.001***	0.001***	0.002***	0.002***	0.001**	0.002***	0.002***
	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0005)	(0.0004)	(0.0005)
U_cool	0.00002		0.00002	0.00001	0.00001	0.00002	0.00001	0.00001
	(0.00001)		(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
B_clust1	-0.189**	-0.184*	-0.163	-0.155	-0.186*	-0.197*	-0.200**	-0.226**
	(0.095)	(0.095)	(0.102)	(0.097)	(0.103)	(0.102)	(0.098)	(0.103)
B_clust10	-0.334***	-0.331***	-0.361***	-0.321***	-0.347***	-0.377***	-0.347***	-0.361***
	(0.086)	(0.086)	(0.091)	(0.088)	(0.092)	(0.092)	(0.089)	(0.092)
B_clust11	0.125	0.129	0.127	0.181	0.133	0.091	0.141	0.090
	(0.140)	(0.140)	(0.149)	(0.143)	(0.151)	(0.150)	(0.145)	(0.151)
B_clust12	0.105	0.107	0.061	0.116	0.050	0.075	0.141*	0.064
	(0.079)	(0.079)	(0.084)	(0.081)	(0.085)	(0.084)	(0.082)	(0.085)
B_clust13	0.024	0.019	0.028	0.016	0.040	0.009	-0.010	0.023
	(0.083)	(0.083)	(0.088)	(0.085)	(0.089)	(0.088)	(0.086)	(0.089)

B_clust14	-0.213**	-0.212**	-0.187*	-0.212**	-0.228**	-0.177*	-0.209**	-0.228**
	(0.095)	(0.095)	(0.100)	(0.097)	(0.101)	(0.101)	(0.099)	(0.102)
B_clust15	-0.363***	-0.360***	-0.336**	-0.370***	-0.379***	-0.312**	-0.348**	-0.363**
	(0.132)	(0.131)	(0.141)	(0.134)	(0.142)	(0.141)	(0.136)	(0.143)
B_clust16	-0.199**	-0.198**	-0.194**	-0.193**	-0.223**	-0.209**	-0.219**	-0.241**
	(0.090)	(0.090)	(0.096)	(0.092)	(0.097)	(0.097)	(0.093)	(0.098)
B_clust17	-0.058	-0.058	-0.052	-0.051	-0.031	-0.083	-0.092	-0.062
	(0.073)	(0.073)	(0.078)	(0.074)	(0.079)	(0.078)	(0.075)	(0.079)
B_clust18	-0.315***	-0.306**	-0.408***	-0.295**	-0.419***	-0.403***	-0.314**	-0.429***
	(0.122)	(0.121)	(0.128)	(0.124)	(0.129)	(0.128)	(0.125)	(0.129)
B_clust19	-0.018	-0.017	-0.033	-0.003	-0.021	0.022	0.047	0.028
	(0.090)	(0.090)	(0.095)	(0.092)	(0.096)	(0.095)	(0.093)	(0.096)
B_clust2	-0.095**	-0.100**	-0.148***	-0.114***	-0.163***	-0.171***	-0.151***	-0.185***
	(0.043)	(0.043)	(0.045)	(0.044)	(0.046)	(0.045)	(0.044)	(0.046)
B_clust20	0.179	0.183	0.314	0.160	0.356	0.317	0.143	0.359
	(0.408)	(0.408)	(0.437)	(0.417)	(0.442)	(0.439)	(0.422)	(0.444)
B_clust21	0.157	0.159	0.201	0.185	0.193	0.255*	0.222*	0.232*
	(0.127)	(0.127)	(0.136)	(0.130)	(0.137)	(0.136)	(0.132)	(0.137)
B_clust22	-0.113	-0.109	-0.020	-0.117	-0.052	0.007	-0.099	-0.028
	(0.139)	(0.139)	(0.145)	(0.142)	(0.147)	(0.146)	(0.144)	(0.148)
B_clust23	-0.125	-0.125	-0.212	-0.086	-0.182	-0.165	-0.056	-0.150
	(0.134)	(0.134)	(0.143)	(0.137)	(0.144)	(0.143)	(0.139)	(0.145)
B_clust24	0.216*	0.214*	0.284**	0.226*	0.281**	0.263**	0.190	0.257**
	(0.120)	(0.120)	(0.127)	(0.123)	(0.129)	(0.128)	(0.124)	(0.129)
B_clust25	-0.145	-0.143	-0.213**	-0.130	-0.273***	-0.228**	-0.165*	-0.297***
	(0.096)	(0.096)	(0.101)	(0.098)	(0.102)	(0.101)	(0.099)	(0.102)
B_clust26	-0.334***	-0.332***	-0.444***	-0.314***	-0.463***	-0.464***	-0.348***	-0.495***
	(0.098)	(0.097)	(0.103)	(0.100)	(0.104)	(0.104)	(0.101)	(0.105)
B_clust27	-0.317**	-0.314**	-0.330**	-0.239	-0.282*	-0.408**	-0.318**	-0.351**
	(0.151)	(0.151)	(0.161)	(0.154)	(0.163)	(0.162)	(0.156)	(0.164)
B_clust28	0.076	0.062	0.030	0.021	0.034	-0.004	-0.020	0.014
	(0.327)	(0.326)	(0.349)	(0.333)	(0.353)	(0.351)	(0.337)	(0.354)
B_clust3	-0.018	-0.018	-0.017	0.013	-0.014	-0.015	0.014	-0.021
	(0.093)	(0.093)	(0.099)	(0.095)	(0.101)	(0.100)	(0.096)	(0.101)
B_clust4	-0.045	-0.046	-0.006	-0.041	0.003	-0.006	-0.056	-0.003
	(0.093)	(0.092)	(0.098)	(0.094)	(0.100)	(0.099)	(0.096)	(0.100)
B_clust5	0.266	0.245	-0.027	0.402	-0.064	-0.144	0.304	-0.191
	(0.383)	(0.383)	(0.410)	(0.390)	(0.415)	(0.412)	(0.396)	(0.416)
B_clust6	-0.007	-0.007	-0.015	-0.001	-0.027	-0.010	-0.010	-0.027
	(0.063)	(0.063)	(0.066)	(0.064)	(0.067)	(0.067)	(0.065)	(0.067)
B_clust7	0.276**	0.274**	0.321**	0.272**	0.319**	0.318**	0.263**	0.312**
	(0.118)	(0.118)	(0.126)	(0.120)	(0.128)	(0.127)	(0.122)	(0.128)
B_clust8	-0.073	-0.069	-0.142*	-0.061	-0.149*	-0.127	-0.057	-0.140*
	(0.075)	(0.075)	(0.079)	(0.077)	(0.080)	(0.079)	(0.077)	(0.080)
B_clust9	0.084	0.079	0.013	0.064	-0.005	0.035	0.068	0.020
	(0.115)	(0.115)	(0.121)	(0.117)	(0.123)	(0.122)	(0.119)	(0.123)
Bad_ServiceY	-0.713***	-0.713***	-0.754***	-0.636***	-0.588***	-0.852***	-0.762***	-0.666***
	(0.032)	(0.031)	(0.031)	(0.030)	(0.032)	(0.030)	(0.030)	(0.032)
Bad_FoodY	-0.368***	-0.372***	-0.192	-0.202*	0.101	-0.278**	-0.293**	0.052
	(0.116)	(0.115)	(0.122)	(0.116)	(0.124)	(0.122)	(0.117)	(0.124)
Bad_AmbitionY	0.505	0.486	0.526	0.516	0.464	0.475	0.500	0.409
	(0.330)	(0.329)	(0.353)	(0.336)	(0.357)	(0.355)	(0.341)	(0.358)
Bad_FacilityY	-0.060		-0.236	0.101	-0.038	-0.328*	0.064	-0.100
	(0.176)		(0.188)	(0.179)	(0.190)	(0.189)	(0.181)	(0.190)
Bad_ValueY	-0.057		0.008	-0.052	0.031	-0.019	-0.041	0.009
	(0.411)		(0.440)	(0.419)	(0.445)	(0.441)	(0.425)	(0.446)
Constant	2.622***	2.289***	2.511***	0.881***	1.872***	2.891***	1.244***	3.087***
	(0.414)	(0.221)	(0.303)	(0.296)	(0.306)	(0.043)	(0.076)	(0.043)
Observations	9,496	9,496	9,675	9,496	9,675	9,675	9,496	9,675
R^2	0.542	0.542	0.472	0.522	0.460	0.467	0.508	0.456
Adjusted R^2	0.539	0.539	0.469	0.519	0.457	0.464	0.506	0.453
Residual Std. Error	1.079 (df = 9432)	1.079 (df = 9440)	1.156 (df = 9626)	1.102 (df = 9447)	1.169 (df = 9626)	1.161 (df = 9626)	1.117 (df = 9447)	1.173 (df = 9626)
F Statistic	177.138*** (df = 63; 9432)	202.793*** (df = 55; 9440)	179.293*** (df = 48; 9626)	214.550*** (df = 48; 9447)	170.663*** (df = 48; 9626)	175.694*** (df = 48; 9626)	203.604*** (df = 48; 9447)	168.207*** (df = 48; 9626)
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 8: Comparison of PLM results

PLM - Results Comparison

	Models Evalation							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
score1	0.069***	0.069***	0.154***			0.170***		
	(0.012)	(0.012)	(0.007)			(0.007)		
score1_very.pos1	0.050	0.050	0.239***			0.263***		
	(0.057)	(0.057)	(0.043)			(0.044)		
score1_very.neg1	0.073	0.073	-0.403***			-0.411***		
	(0.063)	(0.063)	(0.044)			(0.045)		
score2	1.943***	1.943***		2.880***			3.063***	
	(0.110)	(0.110)		(0.097)			(0.099)	
score2_very.pos1	0.206***	0.206***		0.056			0.055	
	(0.047)	(0.047)		(0.043)			(0.044)	
score2_very.neg1	0.173***	0.173***		0.221***			0.233***	
	(0.060)	(0.060)		(0.053)			(0.055)	
score3	-0.024	-0.024			0.432***			0.491***
	(0.035)	(0.035)			(0.020)			(0.020)
score3_very.pos1	0.141**	0.141**			0.432***			0.448***
	(0.060)	(0.060)			(0.041)			(0.042)
score3_very.neg1	-0.162***	-0.162***			-0.277***			-0.246***
	(0.057)	(0.057)			(0.046)			(0.046)
R_LDA_topic2	-0.050	-0.050				-0.166***	-0.169***	-0.181***
	(0.069)	(0.069)				(0.059)	(0.057)	(0.058)
R_LDA_topic3	0.017	0.017				0.625***	0.510***	0.542***
	(0.064)	(0.064)				(0.053)	(0.052)	(0.053)
R_LDA_topic4	-0.296***	-0.296***				-0.299***	-0.253***	-0.298***
	(0.071)	(0.071)				(0.059)	(0.058)	(0.059)
R_LDA_topic5	-0.352***	-0.352***				-0.525***	-0.477***	-0.489***
	(0.070)	(0.070)				(0.059)	(0.057)	(0.059)
R_LDA_topic6	-0.144**	-0.144**				-0.510***	-0.462***	-0.477***
	(0.070)	(0.070)				(0.055)	(0.054)	(0.055)
R_LDA_topic7	-0.056	-0.056				0.148***	0.206***	0.171***
	(0.070)	(0.070)				(0.055)	(0.053)	(0.055)
R_LDA_topic8	-0.242***	-0.242***				-0.554***	-0.487***	-0.498***
	(0.073)	(0.073)				(0.062)	(0.060)	(0.062)
R_LDA_topic9	-0.286***	-0.286***				-0.762***	-0.644***	-0.634***
	(0.074)	(0.074)				(0.060)	(0.058)	(0.060)
R_LDA_topic10	-0.324***	-0.324***				-0.393***	-0.295***	-0.443***
	(0.070)	(0.070)				(0.057)	(0.055)	(0.056)
R_Topic1_prob	-0.472	-0.472	1.017**	1.459***	2.093***			
	(0.638)	(0.638)	(0.510)	(0.495)	(0.509)			
R_Topic2_prob	-3.030***	-3.030***	-1.687***	-1.479***	-0.814			
	(0.630)	(0.630)	(0.502)	(0.479)	(0.500)			
R_Topic3_prob	6.709***	6.709***	11.091***	9.969***	10.726***			
	(0.636)	(0.636)	(0.479)	(0.455)	(0.480)			
R_Topic4_prob	0.473	0.473	0.775	0.750*	1.474***			
	(0.612)	(0.612)	(0.478)	(0.453)	(0.476)			
R_Topic5_prob	-1.288**	-1.288**	-1.315***	-1.611***	-0.250			
	(0.626)	(0.626)	(0.486)	(0.462)	(0.487)			
R_Topic6_prob	-2.947***	-2.947***	-1.948***	-1.886***	-1.075***			
	(0.554)	(0.554)	(0.404)	(0.385)	(0.403)			
R_Topic7_prob	1.319**	1.319**	2.698***	3.356***	3.730***			
	(0.595)	(0.595)	(0.432)	(0.409)	(0.429)			
R_Topic8_prob	-2.003***	-2.003***	-1.569***	-1.500***	-0.293			
	(0.645)	(0.645)	(0.511)	(0.488)	(0.512)			
R_Topic9_prob	-3.194***	-3.194***	-2.975***	-2.637***	-1.158**			
	(0.621)	(0.621)	(0.464)	(0.442)	(0.464)			
R_Bad_ServiceY	-0.047*	-0.047*	-0.064**	-0.049*	-0.062**	-0.058**	-0.044	-0.055*
	(0.027)	(0.027)	(0.029)	(0.028)	(0.029)	(0.029)	(0.028)	(0.029)
R_Bad_FoodY	0.001	0.001	-0.019	-0.011	-0.030	0.008	0.020	-0.002
	(0.118)	(0.118)	(0.125)	(0.120)	(0.125)	(0.127)	(0.123)	(0.126)
R_Bad_AmbienceY	-0.007	-0.007	0.055	0.021	0.054	-0.163	-0.172	-0.162
	(0.340)	(0.340)	(0.364)	(0.345)	(0.364)	(0.370)	(0.354)	(0.367)
R_Bad_FacilityY	-0.010	-0.010	0.124	-0.005	0.169	0.125	-0.017	0.173
	(0.188)	(0.188)	(0.194)	(0.191)	(0.193)	(0.197)	(0.196)	(0.195)
R_Bad_ValueY	-0.103	-0.103	-0.055	-0.009	-0.064	-0.281	-0.191	-0.256
	(0.455)	(0.455)	(0.452)	(0.462)	(0.451)	(0.459)	(0.474)	(0.455)
U_useful	-0.00001	-0.00001	-0.00002	-0.00000	-0.00001	-0.00002	-0.00000	-0.00002
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
U_fans	0.001***	0.001***	0.001**	0.002***	0.002***	0.001	0.002***	0.001***
	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0005)

U_cool	0.00002 (0.00001)	0.00002 (0.00001)	0.00002 (0.00001)	0.00001 (0.00001)	0.00001 (0.00001)	0.00002 (0.00001)	0.00001 (0.00001)	0.00001 (0.00001)
Constant	1.934*** (0.422)	1.934*** (0.422)	1.915*** (0.308)	0.307 (0.299)	1.299*** (0.308)	2.690*** (0.048)	0.954*** (0.079)	2.920*** (0.048)
Observations	9,496	9,496	9,675	9,496	9,675	9,675	9,496	9,675
R^2	0.592	0.592	0.528	0.560	0.530	0.511	0.553	0.521
Adjusted R^2	0.590	0.590	0.527	0.559	0.529	0.510	0.552	0.520
F Statistic	391.480*** (df = 35; 9460)	391.480*** (df = 35; 9460)	538.861*** (df = 20; 9654)	603.224*** (df = 20; 9475)	543.430*** (df = 20; 9654)	503.620*** (df = 20; 9654)	585.041*** (df = 20; 9475)	525.055*** (df = 20; 9654)
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 9: Comparison of PLM results with best imputation approach

Comparison of Imputation Methods of Score2 - PLM

	Evaluation of Imputation Methods of Score2 - PLM				
	(1)	(2)	(3)	(4)	(5)
score2	2.138***	2.831***	2.833***	2.812***	2.521***
	(0.079)	(0.097)	(0.097)	(0.097)	(0.091)
score2_very.pos1	0.272***	0.047	0.046	0.055	0.142***
	(0.040)	(0.043)	(0.043)	(0.043)	(0.042)
score2_very.neg1	-0.148***	-0.168***	-0.169***	-0.163***	-0.133**
	(0.045)	(0.053)	(0.053)	(0.053)	(0.053)
R_Topic1_prob	1.642***	1.381***	1.371***	1.566***	1.657***
	(0.489)	(0.485)	(0.485)	(0.485)	(0.490)
R_Topic2_prob	-1.444***	-1.439***	-1.441***	-1.456***	-1.474***
	(0.482)	(0.478)	(0.478)	(0.478)	(0.483)
R_Topic3_prob	10.145***	10.070***	10.067***	10.017***	10.116***
	(0.458)	(0.455)	(0.455)	(0.455)	(0.460)
R_Topic4_prob	0.810*	0.693	0.687	0.698	0.884*
	(0.455)	(0.451)	(0.451)	(0.452)	(0.458)
R_Topic5_prob	-1.749***	-1.544***	-1.545***	-1.532***	-1.836***
	(0.464)	(0.461)	(0.461)	(0.461)	(0.465)
R_Topic6_prob	-1.896***	-1.990***	-1.991***	-1.976***	-1.905***
	(0.386)	(0.384)	(0.384)	(0.384)	(0.387)
R_Topic7_prob	3.427***	3.412***	3.409***	3.374***	3.302***
	(0.412)	(0.407)	(0.407)	(0.408)	(0.414)
R_Topic8_prob	-1.629***	-1.494***	-1.496***	-1.506***	-1.666***
	(0.490)	(0.487)	(0.487)	(0.487)	(0.491)

R_Topic9_prob	-2.772***	-2.686***	-2.690***	-2.689***	-2.785***
	(0.444)	(0.441)	(0.441)	(0.441)	(0.445)
R_Bad_ServiceY	-0.046*	-0.045	-0.045	-0.046*	-0.047*
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
R_Bad_FoodY	0.010	-0.003	-0.004	-0.001	0.016
	(0.120)	(0.119)	(0.119)	(0.119)	(0.120)
R_Bad_AmbienceY	-0.009	0.007	0.006	0.010	-0.028
	(0.349)	(0.347)	(0.347)	(0.347)	(0.350)
R_Bad_FacilityY	0.216	0.125	0.132	0.124	0.297
	(0.185)	(0.184)	(0.184)	(0.184)	(0.186)
R_Bad_ValueY	0.011	-0.127	-0.143	-0.156	-0.129
	(0.433)	(0.430)	(0.430)	(0.430)	(0.434)
U_useful	-0.00000	-0.00000	-0.00000	-0.00000	0.00000
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
U_fans	0.002***	0.002***	0.002***	0.002***	0.002***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
U_cool	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Constant	0.787***	0.518*	0.521*	0.507*	0.660**
	(0.298)	(0.293)	(0.293)	(0.293)	(0.296)
Observations	9,675	9,675	9,675	9,675	9,675
R^2	0.506	0.655	0.662	0.517	0.562
Adjusted R^2	0.505	0.654	0.661	0.516	0.561
F Statistic (df = 20; 9654)	494.840***	916.164***	945.004***	516.558***	619.295***
Note:					*p<0.1; **p<0.05; ***p<0.01

5. Performance Improvement

5.1 Main Model

5.1.1 Imputation for Missing Values

The five imputations methods for the variable “score2” are as follow:

Table 10 Missing Value Handling – Imputation Methods

Method	Description
1	Impute by zero, which is also the approximately the mean of *Score2*
2	Impute by yearly average of “score2” across all observations
3	Impute by monthly average of “score2” across all observations
4	Predict *Score2* using other independent variables (decision tree)
5	Predict *Score2* using other independent variables (PLM)

Since the values of the binary variables “score2_very.pos” and “score2_very.neg” depended on the quantile of “score2” of the dataset, the two values are re-calculated instead of imputed. Next, each of the five imputations and the associated re-calculations are applied on the best (defined as the highest R-squared from the baseline results) OLS and the best (same definition as above) PLM models. The imputation method that generated the highest R-squared would be adopted.

Table 11 Comparison of imputations Methods of Score 2

	OLS Model –R^2	PLM Model–R^2
Imputation 1	0.486	0.506
Imputation 2	0.492	0.655
Imputation 3	0.492	0.662
Imputation 4	0.492	0.517
Imputation 5	0.483	0.562

It is obvious that missing value handling with imputation method 3 generates the largest R^2 as 0.492 for the best OLS model and largest R^2 as 0.662 for the best PLM model. After imputation of those missing value with monthly average, the R^2 of OLS model is 0.492, that is slightly worsen from the baseline OLS model R^2 of 0.496. The R^2 of PLM model is 0.662 that is improved from the baseline PLM model R^2 of 0.559.

5.1.2 Using Different Models for Better Interpretation

Rpart

For easier interpretation, we use Rpart to partition the best model found in stage 4. Result in **Figure 9** shows that sentimental score 2 has major determining power on star rating. If the sentimental score is less than 0.67, the star rating would not exceed 3. If the sentimental score is below 0.48, the rating given would be as low as 1.5. Most of topics mentioned when sentimental score is lower than 0.67 are 4,5,6,8,9,10. On the other hand, when sentimental score is more than 0.67, the frequent topic mentioned is topic 3 and the rating in general is relatively high. **Figure 10** shows the ranking of importance variables. It is obvious that score 2 and topic 3 probability score are those important variables in decision tree.

Figure 9: Rpart Plot on the Main Model

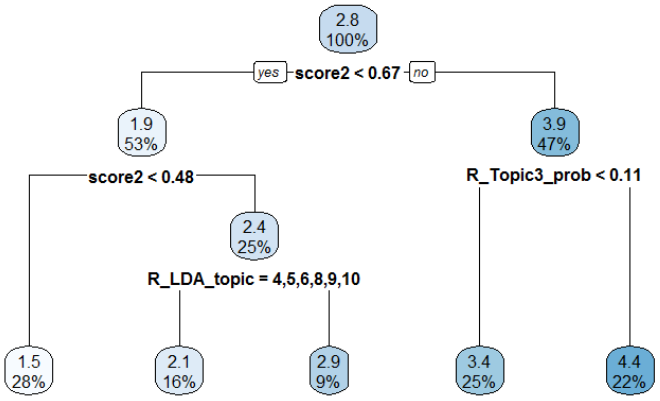


Figure 10: Variable Importance Shown in Rpart

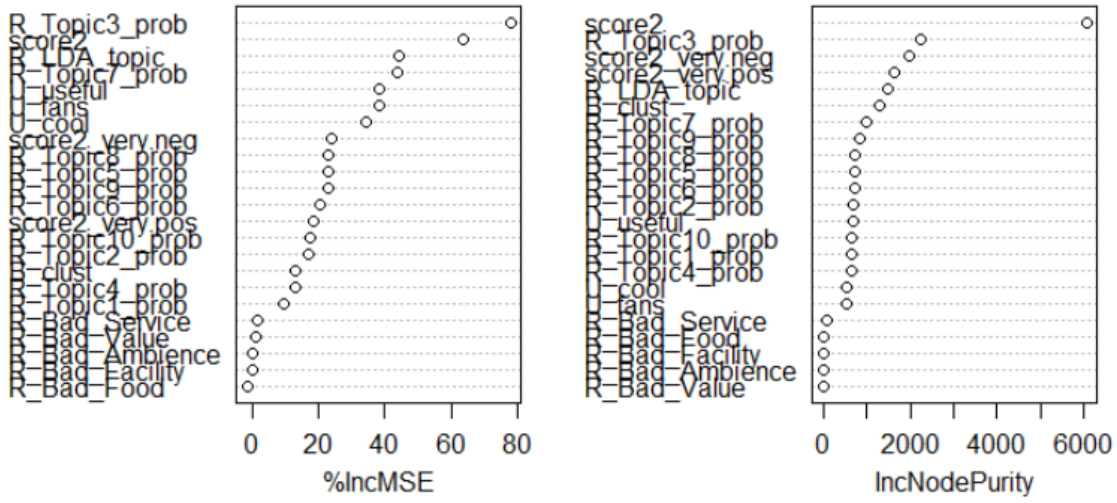
score2	score2_very.pos	score2_very.neg	R_Topic3_prob	R_LDA_topic
10090.96545	5257.13418	4340.10350	3589.69700	3109.64536
U_cool	R_Topic7_prob	R_Topic9_prob	U_fans	R_Topic2_prob
1169.36367	97.90174	87.48875	83.43347	71.23590
R_Topic6_prob	R_Topic10_prob	R_Topic1_prob	R_Topic4_prob	
71.17489	62.57658	56.76016	53.50059	

Random Forest

Random Forest (RF) is used on best model obtained on stage 4. It shows that score 2 and topic 3 are the most important variables contributing to star rating. This is consistent with results obtained in Rpart. Random Forest also provides a mean mse of 1.151469. **Figure 11** shows the variable importance plot obtained from Random Forest. Again, the importance variables are score 2, topic_3_probalility score, which are consistent with findings in Rpart model.

Figure 11: Variable Importance Plot from Random Forest

```
> importance(mod.rf)
                   %IncMSE IncNodePurity
score2          63.93829912  6063.9750676
score2_very.pos 18.60399371  1643.3339003
score2_very.neg 23.89338613  1978.4759409
R_LDA_topic     44.47018654  1481.5425729
R_Topic1_prob    9.69093387   663.6820562
R_Topic2_prob   17.04847718   674.7255680
R_Topic3_prob   78.11571766  2237.1000853
R_Topic4_prob   12.95602384   646.0558864
R_Topic5_prob   22.77539242   737.1356281
R_Topic6_prob   20.59284022   723.5905673
R_Topic7_prob   43.78956474   982.0975956
R_Topic8_prob   22.99534671   742.5413140
R_Topic9_prob   22.71809050   827.3865387
R_Topic10_prob  17.49926368   664.7815901
R_Bad_Service    1.44782947    88.7982639
R_Bad_Food      -1.65743463    11.1917436
R_Bad_Ambience   0.03170198     0.7671337
R_Bad_Facility  -0.26064321     3.2988976
R_Bad_Value      1.00100150     0.1107193
U_useful        38.42985741   674.2085032
U_fans          38.13990055   542.5357006
U_cool          34.50204692   544.7913820
B_clust         13.15190240  1292.7915999
```



GBM

GBM or boosting is used to check the best model and compare the mean square error (mse) with what has been obtained from Random Forest. The mse obtained from boosting is 1.142144, which close to mse of Random Forest 1.151469. This shows that the robustness of best model.

5.2 Other Improvements

5.2.1 Sentiment Score

Details of **Formula 1** and **Formula 3** are as follows:

In **Formula 1,** the sentiment score was calculated by taking the difference of total numbers of matched-positive and matched-negative:

$$score1_i = \text{positive.matches}_i - \text{negative.matches}_i \quad \text{where } i \in \{review\_id\}$$

In **Formula 3**, the numbers of counts of positive (negative) words are first adjusted as a percentage of the average numbers of positive (negative) words appeared in all review pertaining to the franchise, so that the "sentiment" was benchmarked against all other reviews made on the relevant franchise:

$$\text{adj.positive.matches}_i = \frac{\text{positive.matches}_i}{\text{ave}(\text{positive.matches}_j)}$$

$$\text{adj.negative.matches}_i = \frac{\text{negative.matches}_i}{\text{ave}(\text{negaitive.matches}_j)}$$

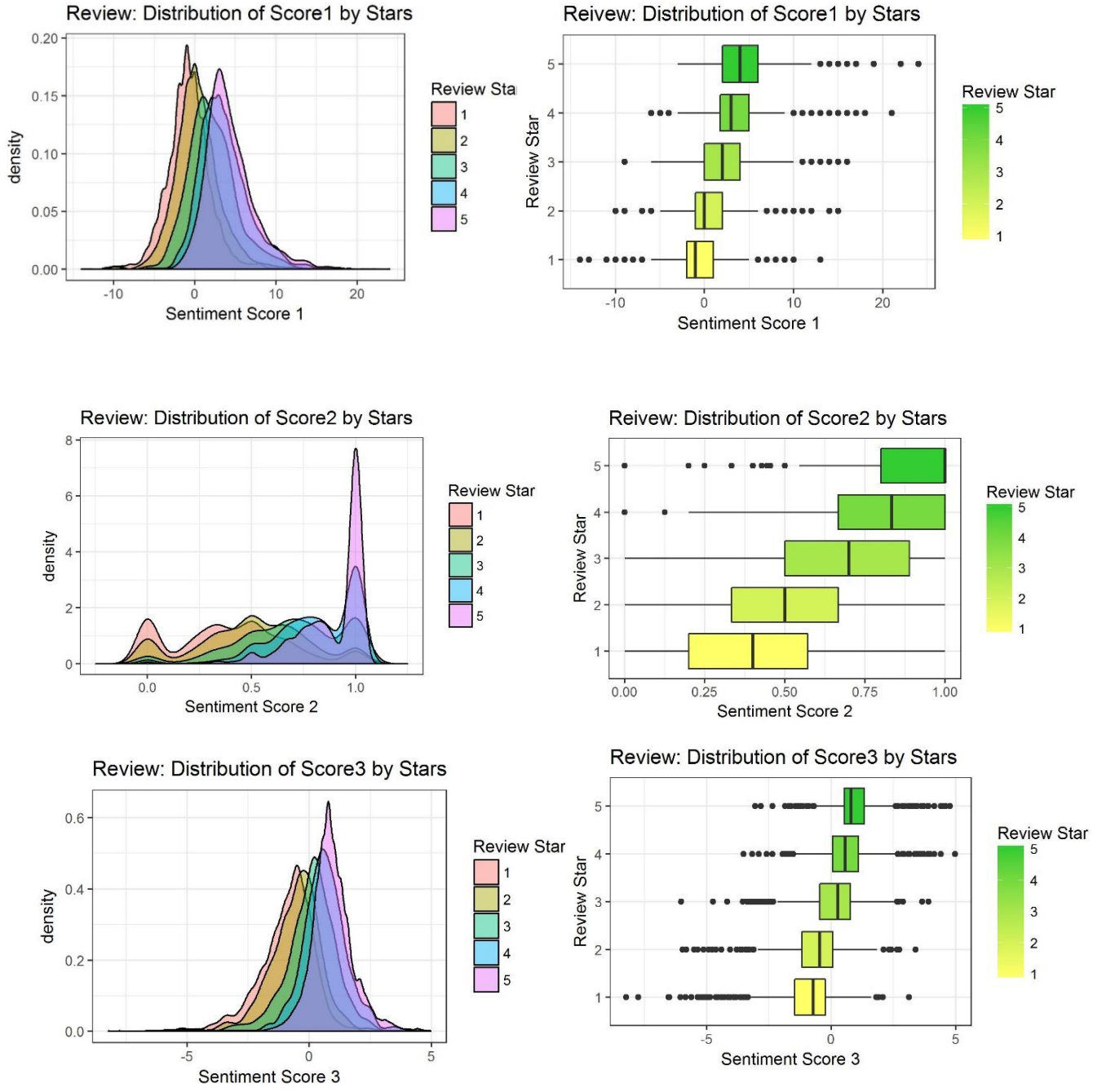
$$score3_i = \text{adj.positive.matches}_i - \text{adj.negative.matches}_i$$

$$\text{where } i \in \{review\_id\},\ j \in \{franchise\}$$

The sentiment scores as calculated by the 3 Formulas are evaluated by the "goodness of fit" to explain review stars.

Figure 12 shows the density and boxplots which visualized the relationship between the sentiment scores (1, 2 &3) with the original *review star* given by the Yelp reviewers. Despite that variation exists in the distribution of sentiment scores (i.e. positive sentiment scores existed for some 1-star or 2-star review while negative sentiment scores existed for some 4-star or 5-star review), there was a clear correlation between the mean of sentiment scores and the review ratings:

Figure 12: Density and Boxplot of Sentiment Scores (by 3 Formulas) against Review Stars



The density distributions of **Score 1 and Score 3** shows a clear parallel shift of sentiment score (from low to high) correspondingly with the review star (from 1 to 5). The boxplot of **Score 2** indicates relatively smaller amounts of "outliers".

To further quantity which score performed the best in "explaining" the review rating, a multi-class regression was done by **Review Star ~ (a combination) of $Score_1$, $Score_2$ and $Score_3$**.

Among models with only one score being used, the model using $\boldsymbol{Score_2}$ produced the lowest residual deviance and AIC and therefore considered as the best fit. Models using more than one score and with $\boldsymbol{Score_2}$ i.e., "2&3" and "1,2&3" gives even lower residual deviance and AIC but the improvement from "2" is not significant.

Table 12: Sentiment Score Evaluation on Reviews (Individual):

Sentiment Score Evaluation - Review Level

Score	Residual.Deviance	AIC
1	20596.29	20612.29
2	19687.26	19703.26
3	20216.48	20232.48
1&2	19417.22	19441.22
1&3	20002.48	20026.48
2&3	19365.19	19389.19
1,2&3	19312.38	19344.38

Robustness 1

Review stars and sentiment scores are further aggregated on the same business id on a weekly basis. By performing a linear regression, $\boldsymbol{Score_2}$ performed consistently the best. Among models using only one of the scores, using $\boldsymbol{Score_2}$ has significant improved R^2 by 0.05; By using more than one score and including $\boldsymbol{Score_2}$ (such as 1&2, 2&3 and 1,2&3), the further improvement in R^2 ranges from 0.02 to 0.03. This is consistent with the observations from above.

Table 13: Sentiment Score Evaluation on Reviews (Aggregated):

Results Comparison

	Review Stars						
	1	2	3	1&2	1&3	2&3	1,2&3
avg_score1	0.256***			0.100***	0.065***		0.019*
	(0.004)			(0.006)	(0.011)		(0.010)
avg_score2		3.279***		2.399***		2.182***	2.168***
		(0.044)		(0.067)		(0.070)	(0.070)
avg_score3			0.760***		0.588***	0.341***	0.294***
			(0.011)		(0.030)	(0.017)	(0.031)
Constant	2.479***	0.847***	2.945***	1.228***	2.827***	1.553***	1.528***
	(0.016)	(0.031)	(0.014)	(0.038)	(0.023)	(0.047)	(0.049)
Observations	7,739	7,552	7,739	7,552	7,739	7,552	7,552
R^2	0.344	0.420	0.372	0.443	0.375	0.449	0.450
Adjusted R^2	0.344	0.420	0.372	0.443	0.375	0.449	0.449
Residual Std. Error	1.226 (df = 7737)	1.156 (df = 7550)	1.200 (df = 7737)	1.133 (df = 7549)	1.197 (df = 7736)	1.126 (df = 7549)	1.126 (df = 7548)
F Statistic	4,051.626*** (df = 1; 7737)	5,472.305*** (df = 1; 7550)	4,575.584*** (df = 1; 7737)	2,999.127*** (df = 2; 7549)	2,317.977*** (df = 2; 7736)	3,079.497*** (df = 2; 7549)	2,054.746*** (df = 3; 7548)

Note: *p<0.1; **p<0.05;** p<0.01

Robustness 2

A 2nd source of dictionary from Harvard General Inquiry Dictionary are used for robustness, which comprise of 1,566 positive words and 2,006 negative words. Compared across the residual deviance, AIC and R^2 from the multinomial and OLS models, dictionary 2 gives a consistent conclusion that $\boldsymbol{Score_2}$ gives the best fit. The fitness however, is slightly worse than dictionary 1. Therefore, the output from dictionary 1 was used as final inputs for the main modelling.

Table 14: Sentiment Score Evaluation on Reviews using Dictionary 2:

Sentiment Score Evaluation - Review Level

Score	Residual.Deviance	AIC
1	21447.67	21463.67
2	21053.10	21069.10
3	21317.04	21333.04
1&2	20733.14	20757.14
1&3	21193.56	21217.56
2&3	20729.92	20753.92
1,2&3	20689.46	20721.46

Robustness 3

Same methodology was also applied on the textual information on *Tip*. As mentioned earlier, since *Tip* does not come with a quantitative rating on individual tip level, the evaluation was performed on aggregated level by business on weekly basis. Across the three formulas, $\boldsymbol{Score_2}$ still provide the best estimation on average review stars on the stores, while the explanation power using textual information from *Tip* is not as good as using those from *Review:*

Table 15: Sentiment Score Evaluation on Tips (Aggregated):

Results Comparison - Tip

	Review Stars						
	1	2	3	1&2	1&3	2&3	1,2&3
avg_score1	0.425***			0.226***	0.076		0.010
	(0.035)			(0.063)	(0.100)		(0.129)
avg_score2		1.678***		0.932***		0.732***	0.732***
		(0.141)		(0.250)		(0.270)	(0.270)
avg_score3			0.269***		0.226***	0.166***	0.161*
			(0.021)		(0.061)	(0.041)	(0.084)
Constant	2.892***	2.016***	3.124***	2.348***	3.083***	2.653***	2.646***
	(0.053)	(0.109)	(0.049)	(0.142)	(0.073)	(0.190)	(0.210)
Observations	840	571	840	571	840	571	571
R^2	0.151	0.200	0.165	0.218	0.165	0.223	0.223
Adjusted R^2	0.150	0.198	0.164	0.215	0.163	0.220	0.218
Residual Std. Error	1.434 (df = 838)	1.420 (df = 569)	1.423 (df = 838)	1.406 (df = 568)	1.423 (df = 837)	1.401 (df = 568)	1.402 (df = 567)
F Statistic	149.571*** (df = 1; 838)	142.047*** (df = 1; 569)	165.297*** (df = 1; 838)	78.955*** (df = 2; 568)	82.899*** (df = 2; 837)	81.310*** (df = 2; 568)	54.114*** (df = 3; 567)

Note: *p<0.1;* ***p<0.05;*** p<0.01

All above are illustrated using the reviews/tips on Starbucks. For Dunkin' Donuts, results are consistent and included in the coding output.

In summary, **Formula 2** applied on *Review Text* on individual level is the best from the above evaluation.

5.2.2 Text Classification

To evaluate the performance from the 3 labelling methods above, the 200 manually labelled observations are split into 2 data set, 41:200 will be training dataset, and 1:40 will be test dataset. 2 types of algorithm will be used: Support vector machine(SVM), and random forest(RF). For SVM, 2 kernels are tried, Linear and Radial. The better performing one will be taken into consideration. The label "Service" is selected for this testing. Here is the testing result:

Table 16: Text Classification: Evaluation Results on 3 Labelling Methods:

Precision	**Support Vector Machine**	**Random forest**
3 Labels (Positive, Negative, Not Mentioned)	0.697	0.497
2 Labels (Positive, Not Positive)	0.565	0.645
2 Labels (Negative, Not Negative)	0.67	0.88

Recall	Support Vector Machine	Random forest
3 Labels (Positive, Negative, Not Mentioned)	0.477	0.513
2 Labels (Positive, Not Positive)	0.565	0.64
2 Labels (Negative, Not Negative)	0.59	0.625
F Score	**Support Vector Machine**	**Random forest**
3 Labels (Positive, Negative, Not Mentioned)	0.49	0.497
2 Labels (Positive, Not Positive)	0.55	0.62
2 Labels (Negative, Not Negative)	0.585	0.63

As we can see, in all 3 measures, Precision, Recall and F Score, 2 labels (Negative, Not Negative) with Random forest performs the best, which will be selected to build the model for all 5 labels, and predict the labels for the rest of the observations.

5.2.3 Topic Modelling

For model evaluation purpose, LDA model is run with parameter k tuning. K is set as optimal k decided by the model, 5 and 10 respectively for each run. Next, for review rating prediction, multinomial logistic regression is applied on the assigned topics and the associated probabilities with each topic assignment. For the LDA model that generated the lowest AIC, the assigned topics or the associated probabilities would be adopted as the new variables for the main model.

Multinomial logistic regression models

Below describes the regressors used for each of the multinomial logistic regression model.

Score_all - Regressors are the associated probabilities of each LDA topic assignment for all review ratings, where optimal k (k=2) is applied for the LDA model

review_stars ~ topic1.associatedprob + topic2.associatedprob

Score_allk5 - Regressors are the associated probabilities of each LDA topic assignment for all review ratings, where k = 5 is applied for the LDA model

review_stars ~ topic1.associatedprob + topic2.associatedprob + topic3.associatedprob + topic4.associatedprob + topic5.associatedprob

Score_allk10 - Regressors are the associated probabilities of each LDA topic assignment for all review ratings, where k = 10 is applied for the LDA model

review_stars ~ topic1.associatedprob + topic2.associatedprob + topic3.associatedprob + … + topic10.associatedprob

Score_all_topic - Regressors are the 2 topics assigned for all review ratings, where optimal k (k=2) is applied

review_stars ~ lda_topic

Score_allk5_topic - Regressors are the 5 topics assigned for all review ratings, where k = 5 is applied

review_stars ~ lda_topic_k5

Score_allk10_topic - Regressors are the 10 topics assigned for all review ratings, where k = 10 is applied

review_stars ~ lda_topic_k10

Topic Modelling Evaluation

Table 17: Associated probabilities with each topic assignment

Topic Modelling Evaluation

Score	Residual.Deviance	AIC
score_all	5013.348	5029.348
score_allk5	4643.740	4683.740
score_allk10	4371.025	4451.025

Table 18: Assigned Topics

Topic Modelling Evaluation - topic

Score	Residual.Deviance	AIC
score_all_topic	5179.867	5195.867
score_allk5_topic	5520.397	5536.397
score_allk10_topic	5612.007	5628.007

Since the associated probabilities with each topic assignment from the LDA model with parameter k = 10 generated the lower AIC, they would be added as the new variables for the main model.

5.2.4 Word Cloud

For the generation of word clouds for Starbucks, unigram, bigram and trigram are all used to evaluate the output. With the increase in numbers of gram used (i.e. numbers of words counted as a unit to determine the frequency), frequency of each outcome decreased but the word clouds provide more detailed and intuitive information about what differentiates 5-star and 1-star reviews.

Specifically, on *Review Text*, **unigram** generates output of words of high frequency such as "drink", "order", "coffee", "locate" in both 5-star and 1-star reviews. In **bigram**, the situation improved by highlighting certain area such as "customer service", "drive through" but doesn't provide the idea whether "service" are good or bad. In **trigram** as shown in Section 2.1, the word clouds give a clear picture of what categorize a review as 5-star and 1 star. On *Tip Text*, however, the results from trigram are too sparse to interpret and therefore **bigram** is adopted.

Figure 13 and **14** shows the output and frequency of words using **unigram** and **bigram** on *Review Text* and **Figure15** shows the frequency of words using **trigram** on *Tip Text*.

Figure 13: Word Cloud on Reviews on Starbucks (Unigram)

5-Star Reviews

1-Star Reviews

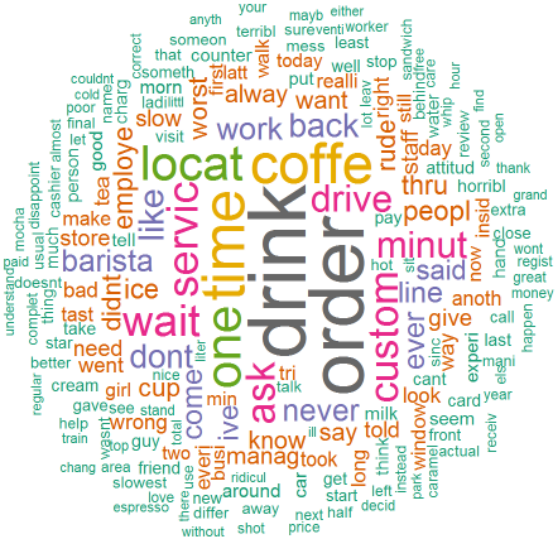


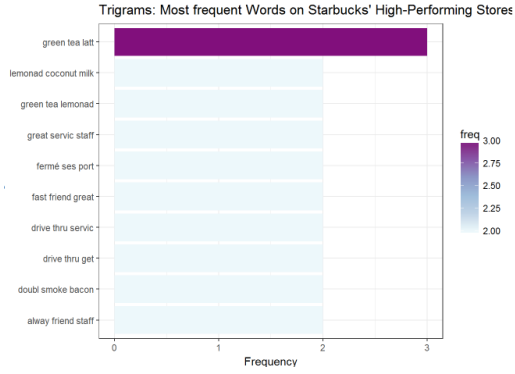
Figure 14: Word Cloud on Reviews on Starbucks (Bigram)

5-Star Reviews

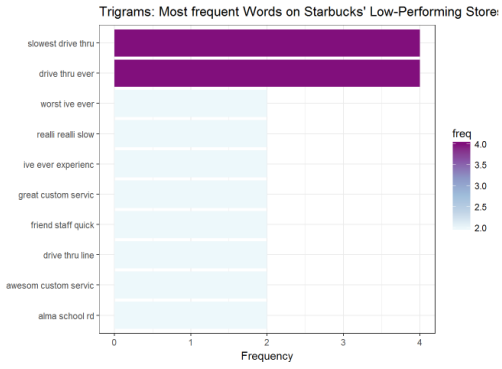
1-Star Reviews

Figure 15: Word Cloud on Tips on Starbucks (Trigram)

High-Performing Stores



Low-Performing Stores



5.2.5 K-means Clustering on User Profile

By including k-means user clustering into the best module, the actual R^2 of the model drops to 0.487. This is shown in **Figure 16**. This enhancement does not improve the performance of the main model, may be due to the curse of dimensionality. Even the best performance with R^2= 0.651 which below the best R^2 of 0.662 shown it **Table 10**. This implies that clustering based on user profile does not improve in review rating.

Figure 16: Partial modeling result with incorporation of K-means clustering of user

U_fans.x	0.001***	0.002***	0.001***
	(0.0004)	(0.0004)	(0.001)
U_cool.x	0.00001	0.00001	0.00001
	(0.00001)	(0.00001)	(0.00001)
as.integer(U_average_stars.x)	0.526***		
	(0.013)		
km.cluster		0.030	
		(0.059)	
kmscale.cluster			-0.007
			(0.057)
Constant	-1.037***	0.442	0.558
	(0.275)	(0.341)	(0.412)
Observations	9,675	9,675	9,675
R^2	0.589	0.487	0.651
Adjusted R^2	0.588	0.486	0.651
F Statistic (df = 21; 9653)	658.840***	435.965***	857.914***
Note:			*p<0.1; **p<0.05; ***p<0.01

6. Conclusion

The purpose of this project is to analyze important factors affecting the user rating in Yelp platform. It is well known that potential customer will refer to Yelp.com to for comments and suggestions when they decide to buy certain product or consume certain service. The rating in Yelp becomes very important for the business, as positive rating would boost business reputations and gives customer assurance and confidence on the product and/or service provided. If the business can figure out what influence those ratings, they will be able to focus more on related marketing effort and influent those impactful factors, and eventually improve the rating and subsequently business reputation. This can be part of Know-Your-Customer (KYC) exercise as well.

In this project, we extracted the data of Starbucks and Dunkin' Donuts in recent 2 years. These are the chain restaurants, which supposed to have standardized product and service. However, the rating varies. It is intriguing to investigate what the factors are affecting those ratings, as it would be a good indication of areas where franchise owners or the chain management can improve. To achieve this purpose, we have set the rating as our dependent variable, the data is constructed at review ID level, and the independent variables are mostly constructed by text mining techniques. From the data, there are some variables we can directly use, such as location, useful, funny and cool. Besides them, we have used various text mining methods to synthesize variables from reviews analytics. The sentimental analysis is used to create sentimental scores; text classification is used to create 5 variables to evaluate the business; topic modelling is used to construct 10 topics from the reviews. During the construction of the additional variables and building the final model on ratings, we have used various algorithms such as OLS, PLM, RPART, Boosting, SVM, Random Forest, LDA, as well as K mean and PCA. Various techniques of handling missing values are also used to improve the final model performance.

From the model result, we can identify quite a few business insights that if follow up well, will lead to better business outcome:

Firstly, sentimental score relates to the rating the most. The more positive sentiment is, higher the rates are. This is expected, as sentiment score reflects customer's feelings from their reviews. Thus, providing an overall good experience, i.e. upkeep service leave and enhance on food, would inevitably lead to happy dining experience on those restaurants.

Secondly, topics on "Always", "Friend", "Fast", "Nice" expectation (topic 4) have the largest positive correlation to rating. This implies that maintain constant staff friendliness, reduce waiting time and create good ambience for customers to hang out with their friends matters a lot to customer's experience. On the other hand, topics on "Ask", "Told", "Look" (topic 7) has the largest negative coefficient. This could be interpreted as communication break down between customers and staffs, and customer is not getting the level of service they expected. Franchise management can invest more on enhancing staff's communication skills and being empathy to customer's request.

Finally, based on binary variables created by text classification, we can find that bad service and bad food adversely impact the rating. The service refers to staff attitude, waiting time, correct order, etc., while food refers to the variety, quality, and freshness of product offered.

In conclusion, with our final model, we can find that to improve rating for chains such as Starbucks and Dunkin' Donuts, it is critical to improve customer's overall experience by providing fast service, being friendly, prepare correct order, provide good quality food, create friendly ambience and good in taking order and communication.

To make the analysis better in the future, we can do more text mining using NLP to extract the text meaning more accurately. Besides, we can use techniques such A/B testing, difference in difference, randomized testing to find out more on the causality on each factor for the customer rating.

Reference

Luca, Michael. 2011. Harvard Business School NOM Unit Working Paper *Reviews ,Reputation ,and Revenue: The Case of Yelp.com*.

Luca, Michael, and G Zervas. 2013. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Harvard Business School NOM …* (April): 1–25.

Monica, Paul R. La. 2017. "Coffee Wars! Wall Street Runs on Dunkin', Not Starbucks - Mar. 23, 2017." *CNN*. http://money.cnn.com/2017/03/23/investing/dunkin-brands-starbucks-coffee-trump/ (April 22, 2017).

Starbucks Corporation. 2016a. "2016 Annual Report of Starbucks Corporation." : 1–150. http://s21.q4cdn.com/369030626/files/doc_financials/2016/Annual/FY16-Annual-Report-on-Form-10-K.pdf (April 20, 2017).

———. 2016b. "Starbucks Corporation - Financial Information - Supplemental Financial Data." http://investor.starbucks.com/financial-information/supplemental-financial-data/default.aspx (April 20, 2017).

Appendix - Index of RMD Code Files and HTML output

Section	File Name	Purpose
N/A	Initial Data Processing	Extract relevant datasets for later processing - All recent 2-year reviews; - Starbucks-related; - Dunkin' Donuts-related;
2.1 & 5.2.4	Exploratory Data Analysis and Word Cloud	- Overview on dependent variable (review star) - Word Cloud
2.2.2 & 3.2.4	Location Clustering	- Create a location categorical variable by clustering
3.2.1 & 5.2.1	Sentiment Score (Dictionary Approach)	- Create a sentiment score variable - Evaluate the results from variations
3.2.2 & 5.2.2	Text Classification	- Create Topic variable by SVM and random forest - Evaluate the results from variations
3.2.3 & 5.2.3	Topic Modelling (LDA)	- Create a LDA Topic variable - Evaluate the results from variations
N/A	Data Merging with New Variables	- Merge datasets with new variables from Section 3
4, 5.1 & 5.2.5	Main Modelling and Enhancement	- Modelling and Enhancement - Imputation of missing values - Create a user group variable by clustering