Student paper
prepared for Prof. Roy Gava

1 Course

Course name:	Quantitative Text Analysis
Semester:	Autumn
Year:	2018

2 Students

Last name:	Yi-Chi	First name:	Chang
Last name:	Hang-Yung	First name:	Hui
Last name:	Wan-Ju	First name:	Yu
Last name:		First name:	

3 Paper

Title:	

Deadline:	10/12/2018	Date:	10/12/2018

1. Abstract

We demonstrate how restaurant reviews can provide useful information for the owner. Our team analyze the reviews of two restaurant: Gordon Ramsay Pub & Grill and Le Village Buffet on Yelp. We adopt text analysis skills such as sentimental analysis and classification and mine out some implications that can help the owner to improve their restaurant. We generate 6 hypothesis relate to the performance of the restaurant. Base on our outcomes, 3 of them are valid,1 hypothesis is invalid but still provide useful information and with 2 hypothesis uncertain. For the implications, our research provides 1. Pros and cons of a restaurant 2. What elements (service, food, atmosphere, etc.) should be taken care of during the festival date 3. What aspect can the restaurant improve their business performance. Overall, our paper demonstrate how catering business can benefit from text analysis.

2. Introduction

Dining is an indispensable part of our daily routine. Instead of cooking the meal at home, people often prefer dining in a restaurant when they are traveling, do not have time to cook or celebrate special occasion. However, to select the right restaurant among a thousand of option is never an easy job. Due to the presence of online platform such as Yelp, Tripadvisor and Google, reviews, ratings and photos have become the critical information for the decision-making process. In addition, comments can provide information about the intangible elements of the restaurant that the images are unable to capture, such as service quality, environment and taste of food. On the other hand, restaurants also rely heavily on reviews and ratings to understand clients' needs and attract new customers. However, given the vast amount of lengthy reviews, restaurants cannot conclude the major feedbacks from clients and identify the way to improve their businesses. Therefore, our team see the opportunity to use text analysis on the restaurant reviews to mine meaningful data for the owner and further interpret the data to generate useful implications or recommendation to restaurant.

3. Data availability

To conduct further analysis, the first step is to check the availability of review data. Our team has tried to acquire review data from Tripadvisor but it does not offer a free API for developer to access to review data. Our team then successfully obtain review data from Yelp. Yelp is currently hosting the Yelp Dataset Challenge that offer review data for personal, educational, and academic purposes. The entire dataset covers 188,593 businesses in the world and 5,996,996 reviews. However, most of the businesses have less than 1,000 review from 2004 - 2018. In order to mine meaningful data, our team purposely filtered the businesses with more than 2,000 reviews. As a result, our team has filtered 74 businesses from the lists with 69 of them are location in Las Vegas. Among the restaurants on the list, our team has selected **Gordon Ramsay Pub & Grill ("GRPG")** and **Le Village Buffet ("LVB")** for the target of analysis.

Gordon Ramsay Pub & Gril: Authentic English club restaurant owned by English chef Gordon Ramsay, who is known by the TV program - Hell's Kitchen, Nosh on elevated British pub food in the comfortable and lively casual bar and lounge seating area, showing sporting events from across the world.

Le Village Buffet: Restaurant in Hotel Paris Las Vegas. The buffet divided into six regions, providing traditional dishes from five different French province. The restaurant is well known for its snow crab legs and Truffle macaroni and cheese. Price range from $25 to $40.

The rationale of selection is that we would like to see the differentiation between a high rating restaurant (Stars > 3) and a low rating restaurant (Stars =< 3). For GRPG, most customers are attracted by the fame of Mr. Gordon Ramsay and may potentially create biases in the review. It could be important for Mr. Gordon Ramsay to understand what his customers like and identify the popular dishes for him to focus on. On the other hand, for LVB, it is more important for the owner to identify the key area to improve the ratings and further attract customers.

4. Definition of scope and hypothesis

To guide the analysis, our team has come up with six initial hypothesis that we would like to explore and mine insights from the review data. We will then analyse and interpret the data to evaluate whether the hypotheses are valid and thus generate recommendations for the restaurant.

Hypothesis	Potential Implications
There are specific elements that most customers will be evaluating at	Advise the business focus on specific areas to enhance customer experience
There are specific elements (e.g. services, food) that are related to low or high rating	Advice the business to put resources for the improvement or presentation of such elements
There are specific problems that will mentioned in majority negative reviews	Advise the business to solve the significant problem
There are specific dishes that mentioned frequently in positive reviews.	Advise the business to put the photo upfront and advice customers what to order
There is high review frequency or good rating in specific holiday.	Advise the business to focus on particular period
There are specific groups of customers being attracted to the restaurant (e.g. couples, families)	Allow the business to know the target customers and offer targeted menu

5. Data collection and preprocessing

To convert the Yelp dataset JSON file that is 4 Gigabytes and contains around 6 millions reviews, our team has applied the `stream_in` function from `jsonlite` package. With reference to the business data from Yelp dataset, we filtered the reviews data by businss_id, “YJ8ljUhLsz6CtT_2ORNFmg” and “ZkGDCVKSdf8m76cnnalL-A” that representing GRPG and LVB respectively. There are 2879 observations for GRPG and 2246 observations for LVB. Samples for the review data is as follow.

review_id	user_id	business_id	stars	date	text
1EwQzhFsHX1C4-Zxs4PwVQ	QNH72vmMZMdyiuaZKh1I8A	YJ8ljUhLsz6CtT_2ORNFmg	5	1/7/2018	So my husband and I came here because we are big Gordon Ramsey fans absolutely love the guy! We had the deviled eggs as our appetizer, I had the fish and chips which was so tender it was falling out the batter and my honey had the lamb burger bomb.com. Our waitress Yaneisy was amazing and the food came out in a reasonable amount of time. Overall good experience looking forward to trying others!
6W3sXdsT3p8rwXjmgnLsqA	KrhohOLwo-ciDTj9qdDv_Q	ZkGDCVKSdf8m76cnnalL-A	2	30/6/2018	Stuff was undercooked or overcooked could've been more variety $89 for two adults and four kids is kind of a lot when the kids don't even eat that much custom made omelette was OK the server making it was probably the best thing at the buffet. Saddened to see that everything is gone away the old Vegas I knew when I was kid is dead

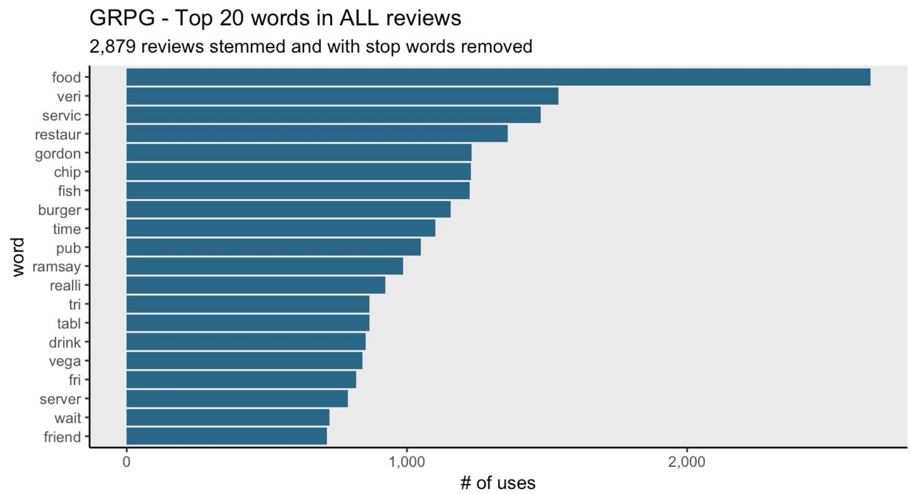
To prepare the review data for further analysis, our team applied the `tm` package to purify the text. Firstly, we used the `tolower()` function to convert all text into lowercase and `str_replace_all()` from `stringr` package to get rid of strange unicode in the text such as é, ñ and <U+00a0>. Then, we removed the punctuation using `removePunctuation()`, removed stopwords using `removeWords()` with the list from `SnowballC` package, stemmed the document using `stemDocument()`. In addition, our team customized the stop words lists beyond `SnowballC` to perform specific analysis and the customized lists will be stated under the Analysis part. For each analysis, our team has generated a data frame containing the subset of the review data and a dataframe that separate the text into individual words (or bigrams and trigrams that will be mentioned specifically). To perform this task, our team used the `unnest_tokens()` function from the `tidytext` package to create a new dataframe from the subset.

6. Analysis and Implications

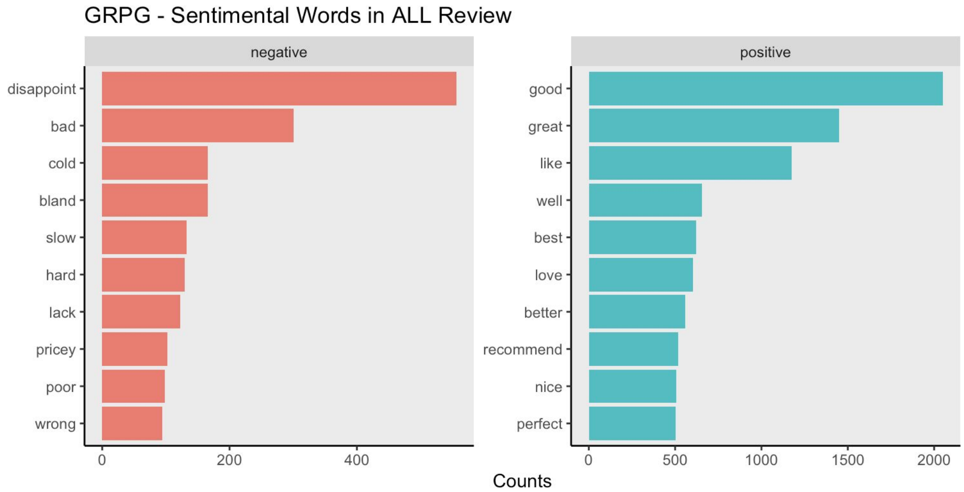
6.1 Opinion Mining

The goal of opinion mining is to **identify the frequently mentioned phrases** in customers' reviews (e.g. disappointing service). The analysis will begin with locating the popular objects which customers are caring the most (e.g. service). Then followed by summarising the commonly used sentimental words (e.g. disappointing) that contribute to high rating reviews (Stars > 3) and low rating reviews (Stars =< 3). Lastly, we will extract bigrams (i.e double words) from the reviews and filter the bigrams based on the objects or the sentimental words to find customers opinions. The frequency will be used to determine the relative importance of the issues.

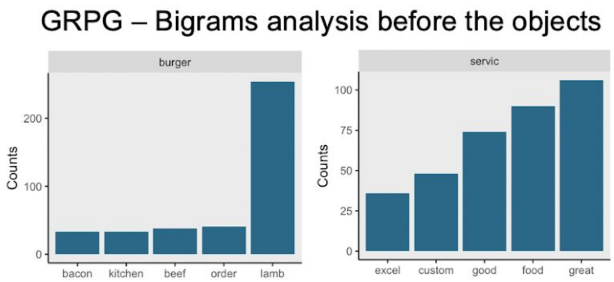
6.1.1 GRPG

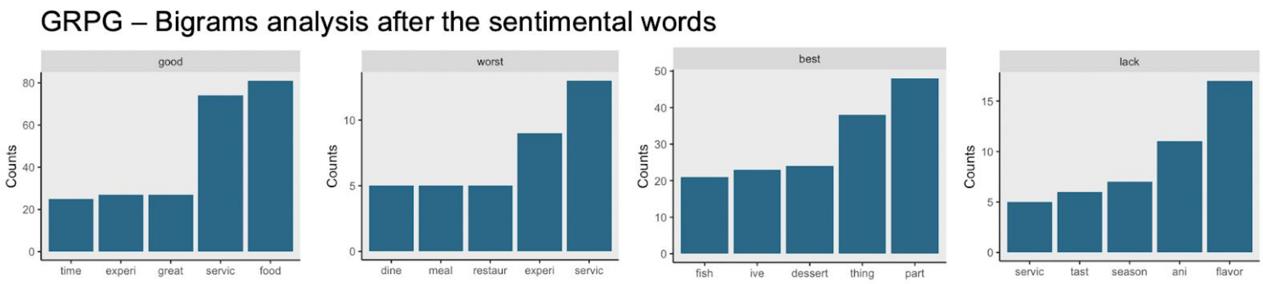


Above is the "objective words" list generated from all GRPG reviews. Our team tried to categorize the words in two buckets, meaningful and meaningless. For meaningful words, it has to be referring several important and specific elements of the restaurants. Examples from the top 20 words are “servic” (service), “chip”, “fish”, “burger”, “time”, “tabl” (table), “drink”, “fri” (fries), “server”, “wait”, “friend”. Reason for several meaningless are as follow. First, “food” is too general for analysis. Second, “veri” (very) and “realli” (really) are adverb that unable to be filtered by `SnowballC` package. Lastly, “gordon”, “ramsay”, “vega” (vegas) and “pub” are referring to the name of the restaurants and locations that is unrelated with our analysis.



Above is the “sentimental words” list generated from all GRPG reviews. Our team then used the `Bing` lexicon to identify the positive and negative sentimental words for all reviews of GRPG. Note that “hell” has been removed from the negative lists as it is classified as negative words but it is actually referring to “hell kitchen” which is one of the TV program hosted by Mr. Gordon Ramsay.





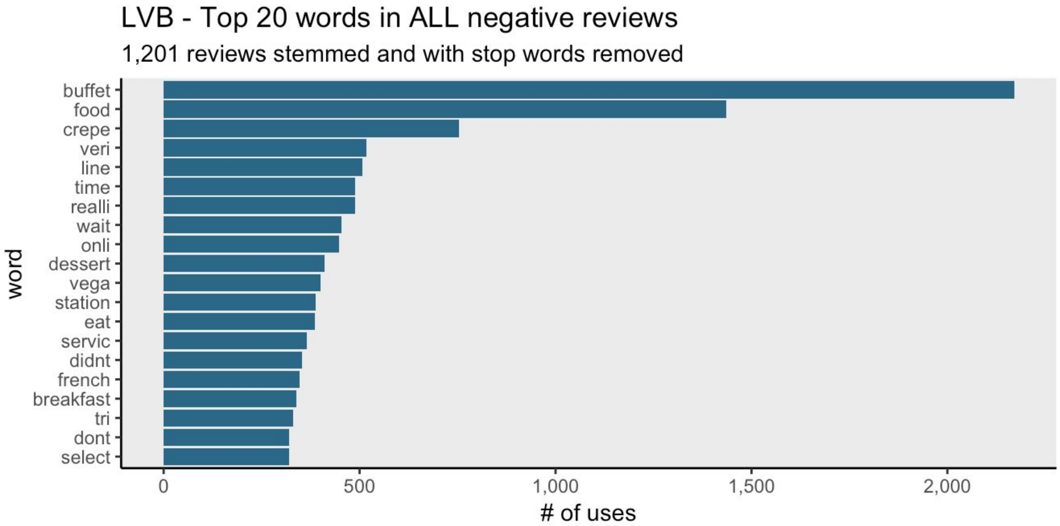
Lastly, we ran the bigrams analysis on the top words from both “objective words” and “sentimental words”. Particularly, our team is interested to see the word prior to the “objective words” so as to understand customers’ opinions to the objects and the word immediately after the “sentimental words” to identify the objects which the emotions are associated with.

For the bigrams analysis on “objective words”, our team observed that the result for several words are very fragmented that is not significant enough to yield implication. “burger” and “servic” are two examples of the meaningful results. For “burger”, the result shows that lamb burger is actually the most popular burger in the restaurant. In fact, we can see from the official website for GRPG that Grilled lamb burger is one of the recommended dishes. The data shows that the recommendation has created a huge impact in terms of clients’ decision making process and attract people to try thought it is no include on the official menu. For “servic”, the number of positive words outperform the negative words which can show that the service level at GRPG is acceptable for majority of customers.

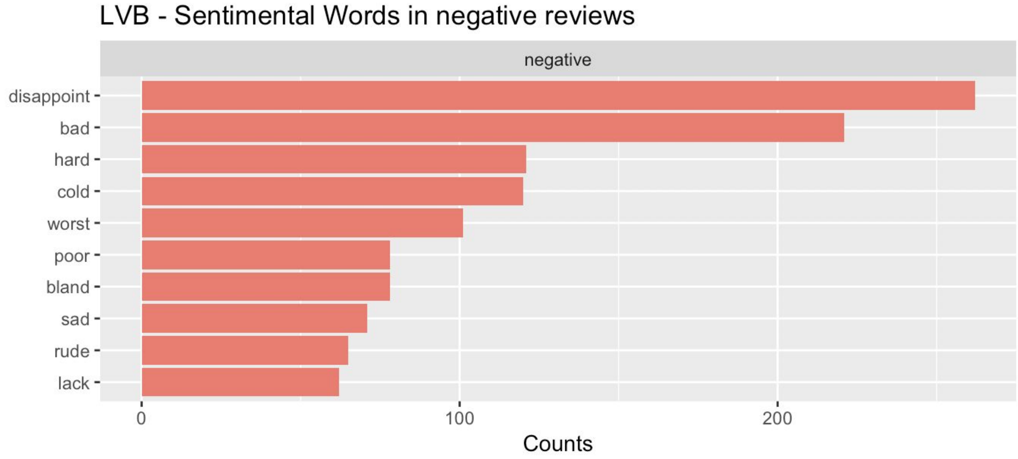
For the bigrams analysis on “sentimental words”, our team presented four example above. Several words actually generated generic result that is not meaningful, such as “good” and “worst” attached above. “disappoint” and “lack” are two examples of the meaningful results. For “best”, our team observed that “dessert” is a significant item that could hint the business to further recommend. For “lack”, most of the customers stated that the flavoring and seasoning could be insufficient. However, our team is unable to further identify the specific dishes.

6.1.2 LVB

While the data for GRPG seems to be one-sided due to the fame of Mr. Gordon Ramsay, our team are interested to explore how the analysis would be for lower-rating restaurant, LVB. To generate insight for LVB to improve their rating, we ran the opinion mining based on all low rating reviews.



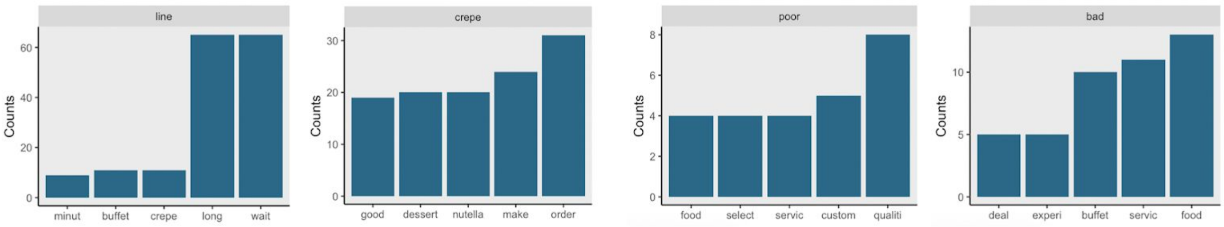
Above is the “objective words” list generated from all LVB low rating reviews. Examples of meaningful words are “crepe”, “line”, “time”, “wait”, “dessert”, “breakfast”. In addition to previous explanation, “onli” (only), “dont” and “didnt” are adverb that unable to be filtered by `SnowballC` package. Then, “buffet” are referring to the name of the restaurants that is unrelated with our analysis.

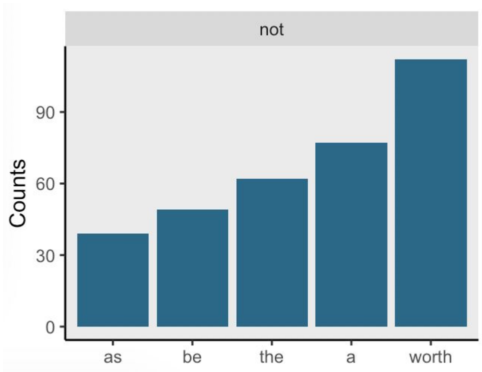


Above is the “sentimental words” list generated from all LVB low rating reviews. The words are similar to that of GRPG.

LVB – Bigrams analysis before the objects

LVB – Bigrams analysis after the sentimental words





After we ran the bigrams analysis, several implications can be drawn from the data. From the “objective words”, we identified that LVB is particularly focus on breakfast and their crepe has been one of the signature dishes which get more than 500 occurrences in 1,201 reviews. The key issue identified with the negative reviews is that there are long line at LVB. With reference to the business model of a buffet restaurant, the line is the queue for getting food from food stations rather than getting a table. In addition, based on the bigram analysis on “sentimental words”, another issue is that the food qualities are not fulfilling the customers’ expectation. Unfortunately, the sample size is relatively smaller and the result is either fragmented or generic so our team decided to use the review data without stop words removed. The output shows that “worth” has more than 100 occurrences immediately after “not”. As a result, the buffet restaurant might be currently perceived to be over-priced by the customers. Business owner could potentially work on the pricing or ensure the food quality to obtain a higher star rating.

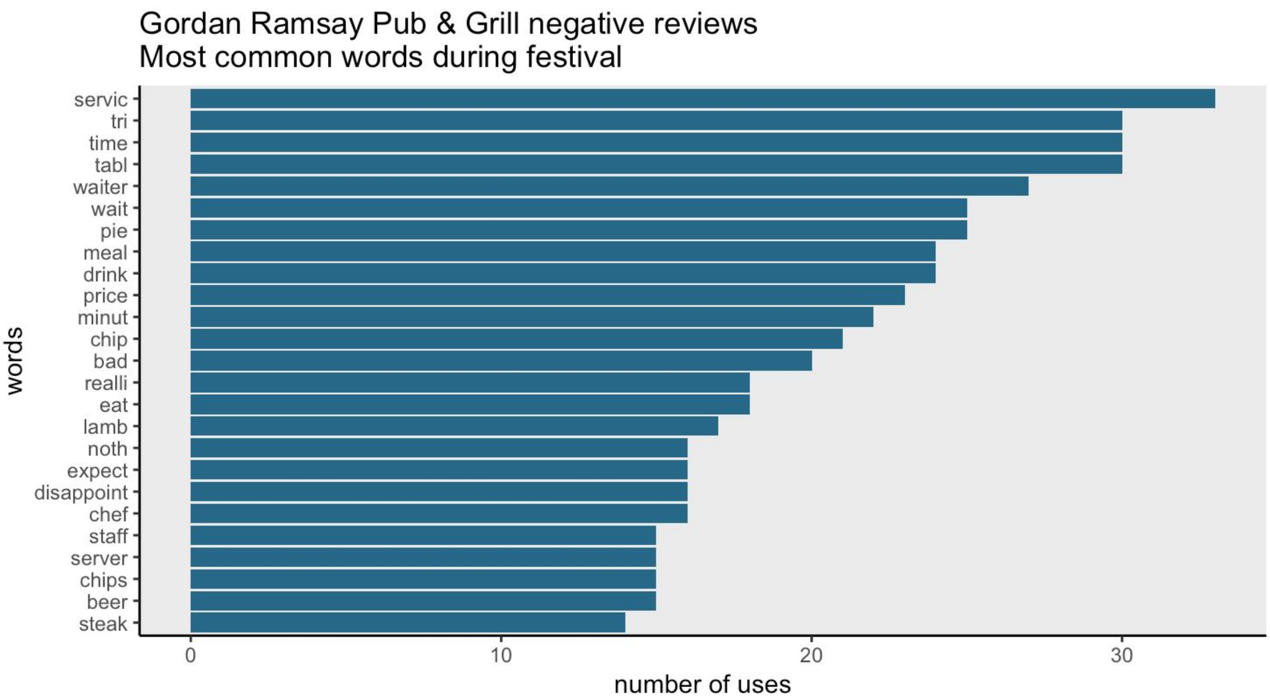
6.2 Time-period analysis (with frequency)

Our purpose of this research is to find out whether customers who went to the restaurant during the festival rate differently from others and why they rate differently. First, In order to find out the difference of the rating behavior, we extract those data that are related to festival date and compare the average rating of the festival data with the mean rating of the original data. Next step, in order to figure out the reason why this difference exist, we extract negative rating review from festival data and tried to find out what kind of words are frequently mentioned in these reviews. With this information, the owner of the restaurant can realizes what kind of element (food, service, etc.) needs to be improved during the festival date.

Comparison: We first extract data that related to specific holidays from the original data by indexing the date of the following holidays: Christmas and New year's eve (12/20~1/1), Valentines day (2/14), Father's day, Mother's day and Thanksgiving (11/24), which are those festivals that people will tend to go out and celebrate in restaurant, and compare the average rating with original data. In order to extract those data, we use `as.Date` function to change the date column in the data into date format so that we can do logical comparison of the date data with our festival date and successfully extract the reviews. Second, in order to simplify the comparison process, we assigned 1 to higher rating reviews and assigned 0 to lower rating reviews. We find out that rating generated during the festival are lower than the average rating of the original data (GRPG 8.3% lower, LVB 6.7% lower).

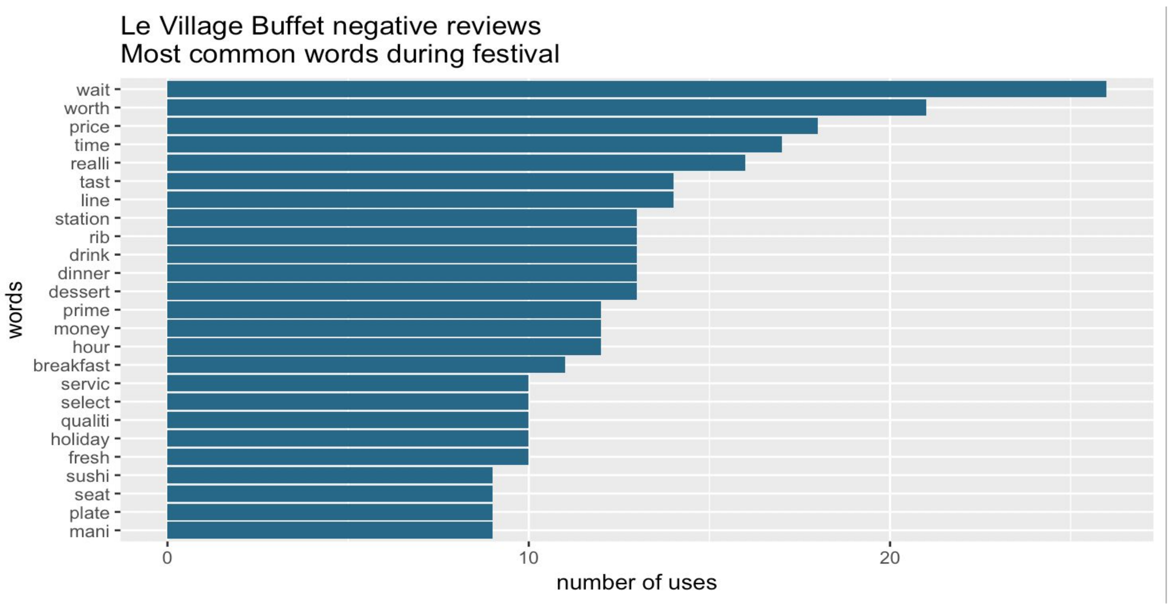
Frequency analysis: To figure out why this difference exist, we extract the lower rating reviews from festival data and try to find out those words that are frequently mentioned in the negative reviews. After removing the stop words mentioned above, we count the number of the remaining words and find out that words like "wait", "line", "time" and "service" are frequently mentioned in the negative reviews in both GRBG and LVB during festival date.

6.2.1 GRPG



We found out that the word “service”, “time”, “table”, “wait”, “waiter” are the meaningful words that are mostly mentioned in the negative reviews during the festival date. With this information, we can probably imply that since the restaurant are too crowd during the festival date, the waiting time gets longer and the quality of the service decrease. The owner of the restaurant can optimize the waiting line to shorten the waiting time and hired more temporary waiter to maintain service quality.

6.2.2 Le Village Buffet



Again, words like “wait”, “time”, “line” ,”hour”, “worth”, “price” are the words that are mentioned in the negative reviews, which can still be implied that the waiting time is too long and lead to a negative rating behaviors.

6.3 Classification

To classify the reviews into different topics so as to analyze which aspects the restaurant owner can improve, we adopt one of the unsupervised machine learning methods, topic modeling, to distinguish between reviews. Since we did not define the category nor label all the reviews in advance, unsupervised machine learning can help us separating all the reviews first, then we can adjust the model according to the results and decide how many potential topics are there in the contents. After deciding the number of topics and obtaining the beta possibility value of each word, we observed words with high beta values to define the concept as well as relevant words to each topic. Thus, the result can be analysed with sentiment analysis to know customers’ attitude towards different topics.

Data Pre-processing for LDA model: Preparation before implementing LDA (Latent Dirichlet Allocation) model includes data preprocessing steps mentioned before, selecting specific stopwords (as below) , creating corpus and create DTM (Document Term Matrix). Used packages are: `readr`, `dyplyr`, `tm`, `stringr`. The method of LDA model is “Gibbs”.

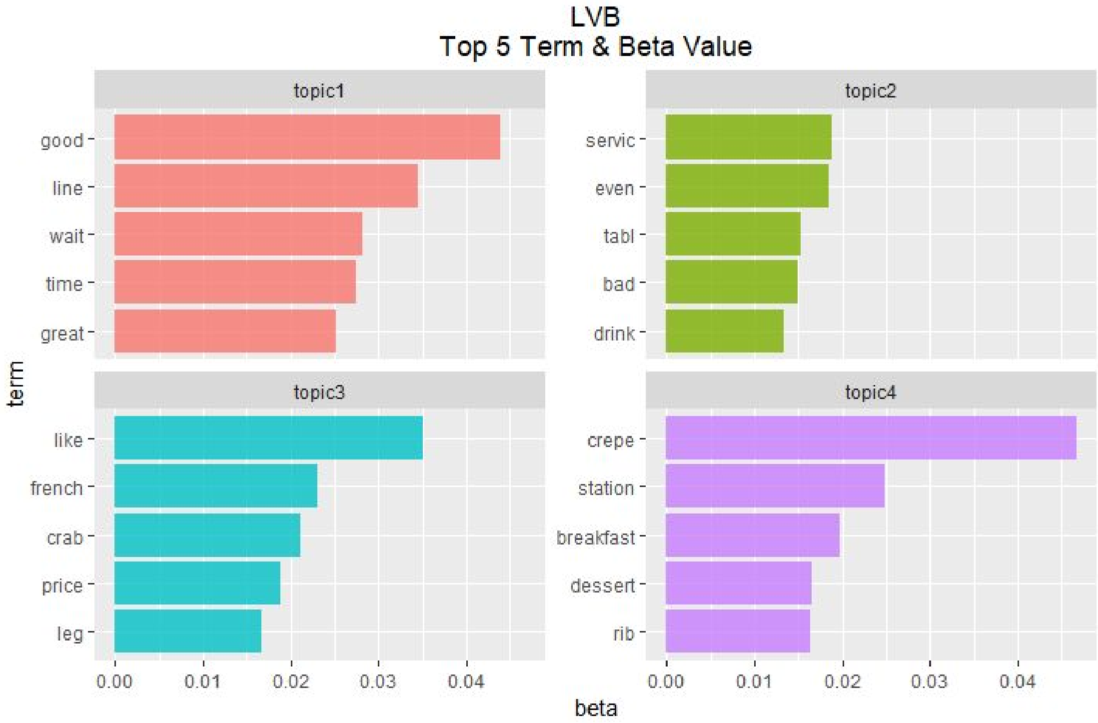
Stopwords of GRPG (extra to `SnowballC` package):

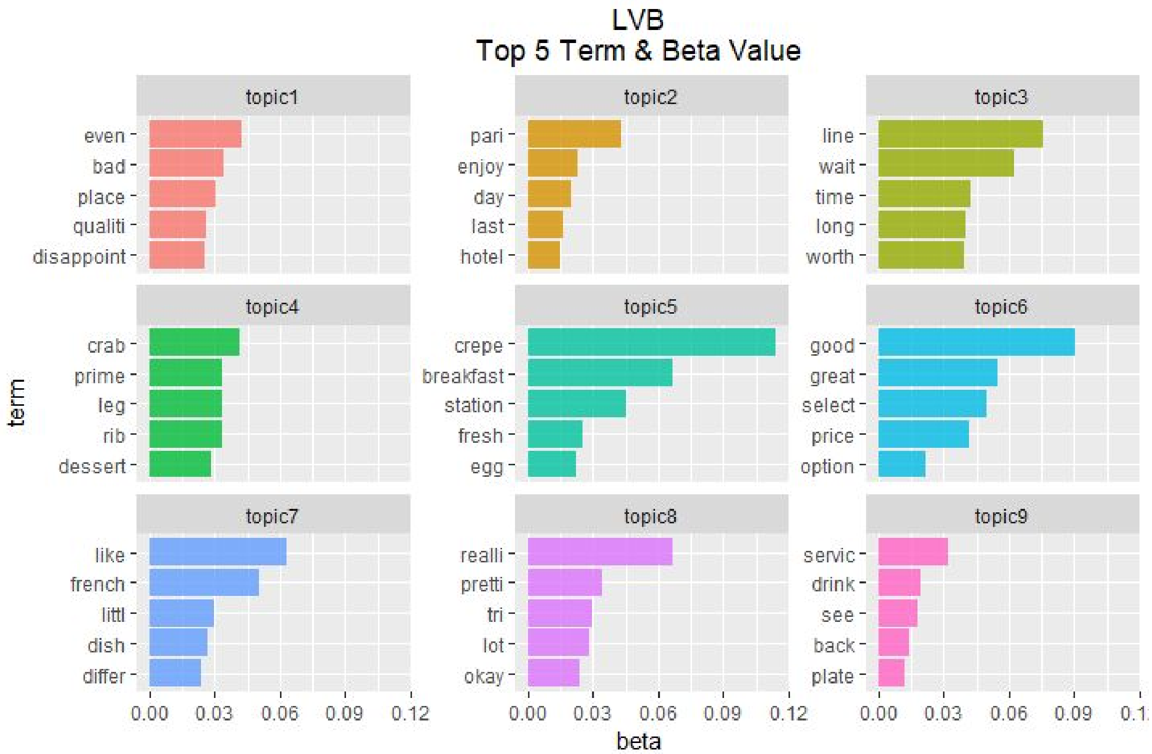
```
c("meal", "eat", 'burger', "food", "gordon", "ramsay", "ramsey","vegas","vega","think",
  "one","two", "three", "four", "five", "six", "seven","eight", "nine", "ten", "just",
  "since","look","restaurant","take","get","say","can","will","feel","wasnt","find","people",
  "person","make","ever","sit","want","visit","though","burgr","hell","ive","didnt","will",
  "know","thing","also","come","much","give", "kitchen")
```

Stopwords of LVB (extra to `SnowballC` package):

```
c("meal", "eat", "food","think","village","vega","vegas","buffet","le","one", "two", "three",
  "four", "five", "six", "seven", "eight", "nine", "ten", "just", "since","look","restaurant",
  "take","get","say","can","will","feel","wasnt","find","people","person","make","ever","sit",
  "want","visit","though","ive","didnt","will","know","thing","also","come","much","give")
```

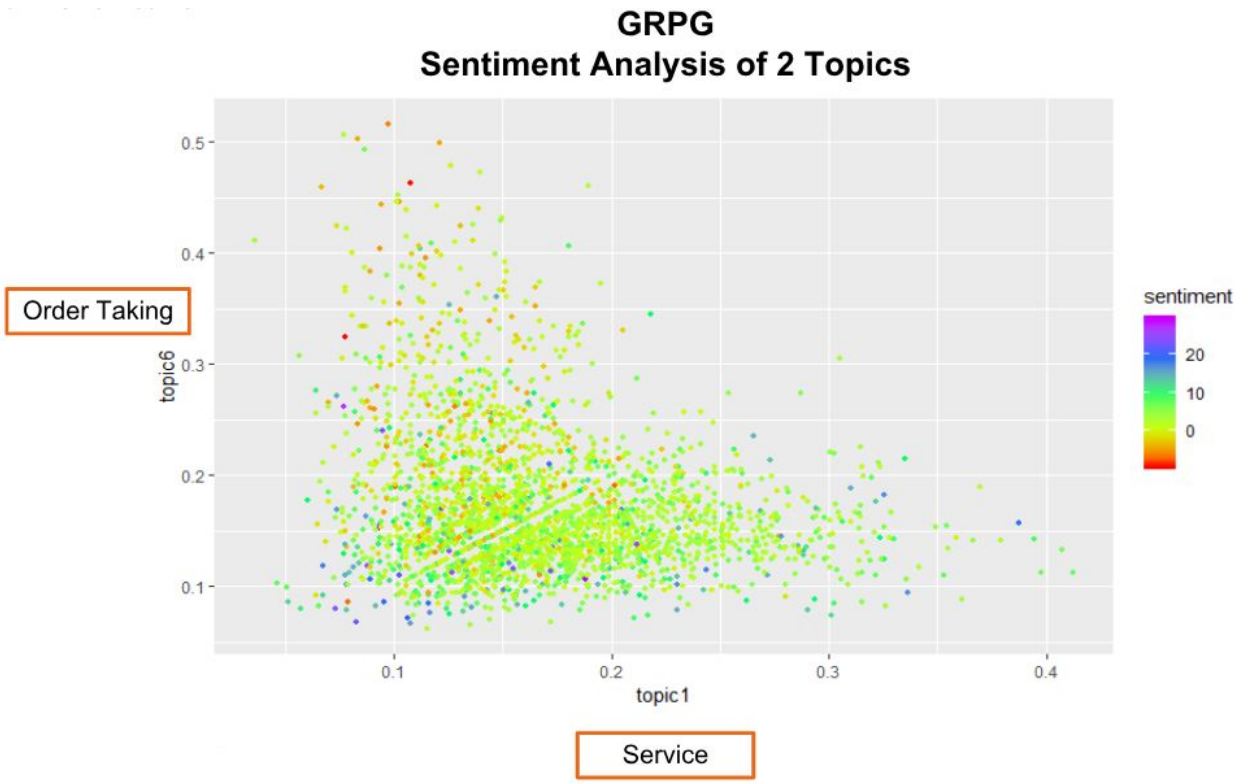
Deciding numbers of topics: Since we did not actually know the underlying topic structure of the reivews, so we tested four, six and nine topics by LDA. As the results displayed below, we compared the words selected and found that, as the topic number n increase, it becomes more specific, however, sacrifices clarity at the same time. Therefore, we chose six as the final optimal topic number.


LVB
Top 5 Term & Beta Value
topic1
good
line
wait
time
great
topic2
servic
even
tabl
bad
drink
topic3
like
french
crab
price
leg
topic4
crepe
station
breakfast
dessert
rib
term
0.00 0.01 0.02 0.03 0.04
beta


LVB
Top 5 Term & Beta Value
topic1
even
bad
place
qualiti
disappoint
topic2
pari
enjoy
day
last
hotel
topic3
line
wait
time
long
worth
topic4
crab
prime
leg
rib
dessert
topic5
crepe
breakfast
station
fresh
egg
topic6
good
great
select
price
option
topic7
like
french
littl
dish
differ
topic8
realli
pretti
tri
lot
okay
topic9
servic
drink
see
back
plate
term
0.00 0.03 0.06 0.09 0.12
beta

6.3.1 GRPG

The graph above displayed top five highest beta value terms of each topic. After the model generated a list of terms with their beta value, we observed the top fifteen of each topic and related those terms with topics from one to six as “service”, “customer expectation”, “food and drinks”, “desserts”, “snacks” and “order taking”.



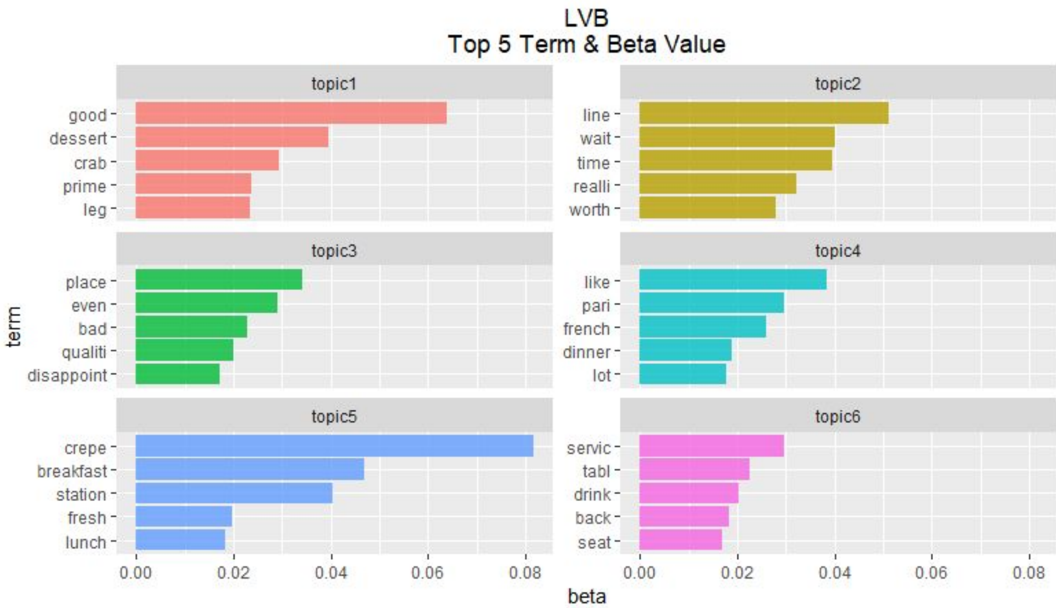
Next step, we would like to see whether customers' reviews reflect issues about these topics, so we chose two of the topics to investigate the relation between the sentiment score and the percentage of mentioning those topics.

Our team calculated the posterior probability and the sentiment score based on AFINN lexicon, then created the document-topic matrix. After filtering the outliers (sentiment score higher than 30, lower than 20), we can draw a scatter plot as above.

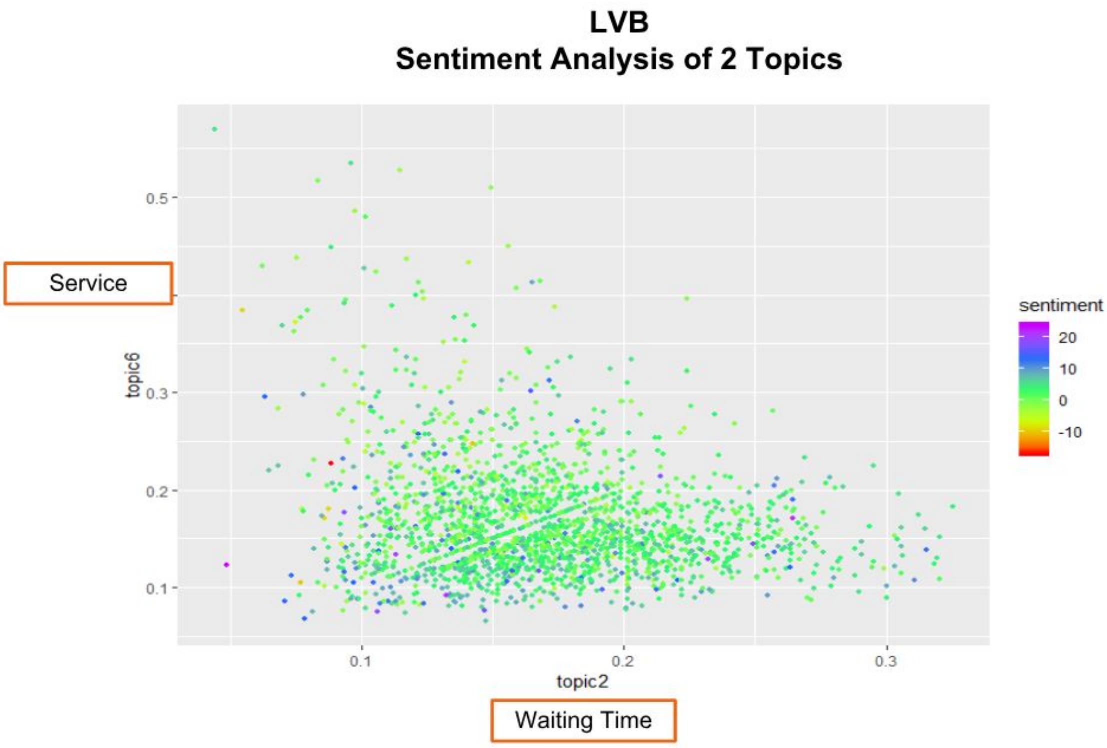
For example, for "service" and "order taking", we can see that reviews with negative sentiment tend to focus more on order taking since reviews marked as red have a higher ratio of order taking / service. In this case, we can conclude that the order taking process of GRPG can be improved while the service provided overall satisfies the customers.

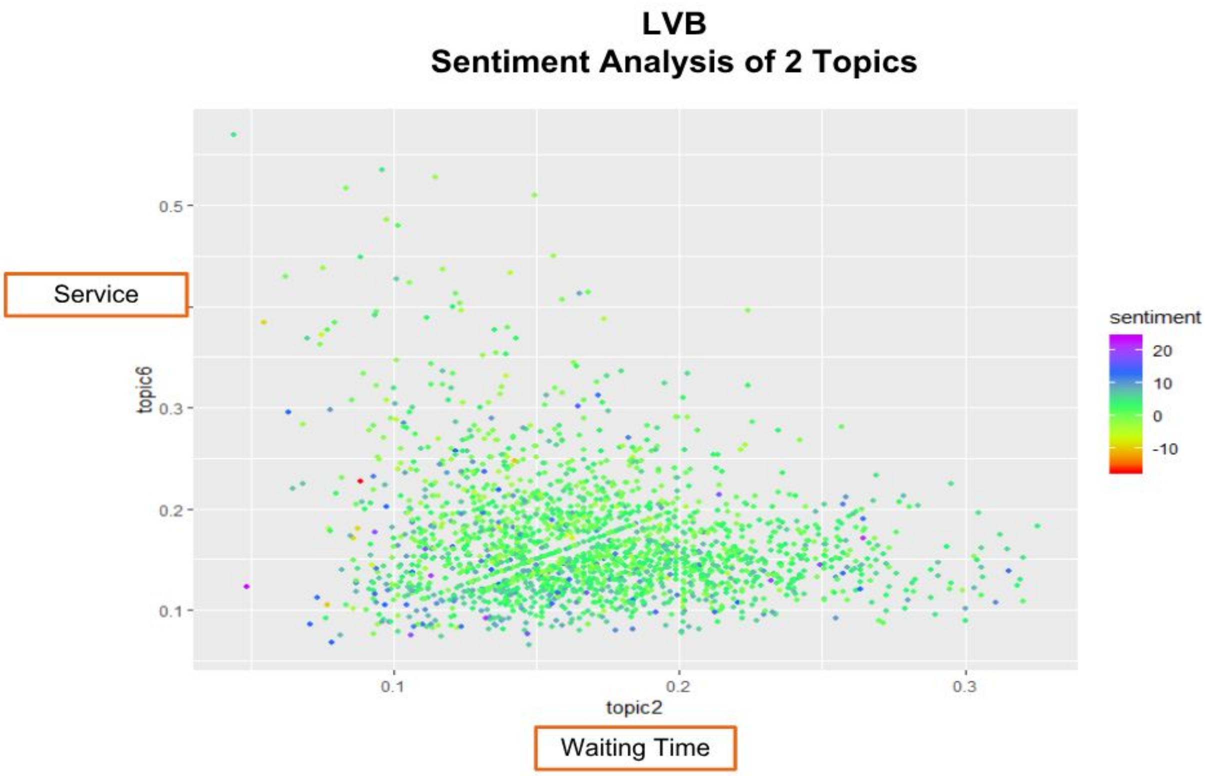
Similar results can be seen by choosing topic 3 "food and drinks" and topic 6 "order taking". Reviews containing negative sentiment tend to mention more about order taking compared to food and drinks in GRPG, which implies that order taking process, in terms of the waiting time and waiter performance, are weak points to GRPG.

6.3.2 LVB



Using the same method, we also obtained six topics from LVB reviews, which are recognized as:"food and desserts", "waiting time", "quality and price", "dinner", "breakfast and lunch" and "service".





For LVB, if we consider the same domain in service and ordering process as GRPG, by observing these two plots above, we can easily notice that people mentioned more on “waiting time” than on ”service”, compared to the one “waiting time” versus “food and desserts”. On the other side, it can also be seen that high sentiment score dots spread evenly on the plot and with only a few negative sentiment scores, meaning that there are no obvious weak points for LVB among food, serving and waiting time.

We can also investigate into other 2 topics of GRPG and LVB, yet we only use service and order taking here as an example.

7. Takeaways and limitations

We believe that the three analyses covered in this paper could be used for other businesses to mine insightful information. The prerequisite is that review data should be available in a large quantities as our team believe that 2,000 reviews are still insufficient to generating useful data. In addition, to apply the above analyses, one should note the following takeaways.

For opinion mining, it is important to try different keywords as the result tends to be generic and fragmented because not all customers would spend time giving detailed reviews and they have different expressions on the same thought. As a result, it is hard to expect that the bigrams analysis could generate a phrase that is highly significant in terms of occurrence. To improve the result, a dictionary could be created to classify scattered results from bigrams analysis into different categories (e.g. “lunch”, “dinner”, “dessert”). The use of categories will allow for a higher significance than the use of individual words. It is much easier to have repeated categories than repeated bigrams. On the other hand, the data should be interpreted with extra information from the restaurant or other platforms. For example, the business model of a restaurant can hint the relevancy of the words and the directions for the implications.

For time-period analysis, we simply use the date of the festival to extract festival data instead of using the reviews that mention the name of the festival(ie. Christmas) since the sample is too small and not robust enough to prove our hypothesis. With this kind of extraction, we can only know what should be improved during the festival “period” but can not focus on what kind of elements are important to the customers who are really celebrating the festival in the restaurant since we also extract reviewers who just visit the restaurant during the festival date with other reasons. With more reviews mentioning specific festival in the text content, we will be able to figure out what kind of elements(atmosphere, price, service...etc.) do the customers cares the most while celebrating the specific festival.

For topic modeling, we found it difficult to apply this model in review classification since the contents of reviews are more highly similar to each other, making it hard to distinguish and generate precise topics. Documents from various subjects may be more easy to implement. Another issue is to decide the numbers of topics. While there are some statistical or mathematical models offering help to choose the optimal number of topics, we did not apply it as it would be useful only after comparing across different models and might not be necessary for our project. In addition, we have tried different topic classification models in the beginning and found that results of using “Gibbs” are much more interpretable than using “VEM”. The main limitations of using topic modeling are that it is very subjective during the process of defining the topics, which means, the different researcher may have different opinions on selecting the main concept from term lists, possibly leading to different conclusions. What’s more, sentiment scores might not perfectly reflect the reviewer’s attitude because the score is related to the length of the paragraph and will be affected by repetition.

8. Conclusion

Based on the analyses in the paper, our team has come to the following conclusion for the 6 hypotheses.

Hypothesis	Results
There are specific elements that most customers will be evaluating at	Valid. The classification analysis shows 6 important topics for each restaurants which are main themes for all reviews.
There are specific elements (e.g. services, food) that are related to low or high rating	Uncertain. The analyses adopted are unable to show a clear difference in elements
There are specific problems that will mentioned in majority negative reviews	Valid. With the example for LVB to focus on queuing in front of food stations.
There are specific dishes that mentioned frequently in positive reviews.	Valid. With lamb burger as the signature dishes for GRPG and crepe for LVB.
There is high review frequency or good rating in specific holiday.	Invalid. Our analysis shows that bad rating are more frequent in both restaurants
There are specific groups of customers being attracted to the restaurant (e.g. couples, families)	Uncertain. The analyses adopted are unable to capture this hypothesis.

Text analysis on review data can bring insights to not only the owner of restaurants but also the customers and the competitors. Restaurants such as LVB can benefit from the text analysis to strengthen their business, attract more customers and improve the star ratings. Customers can benefit from the text analysis to make good decisions and find a lovely restaurant for their special occasion. Competitors can learn the strengths of the restaurant , improve their businesses and be more competitive. Therefore, our team strongly believe that reviews are valuable assets that allow restaurants to generate meaningful information and to enhance their business.

9. Reference

Gordon Ramsay P&G

https://www.caesars.com/caesars-palace/restaurants/gordon-ramsay-pub-and-grill#.XAz-yKfxJQI

https://www.caesars.com/content/dam/clv/Dining/Upscale/gordon-ramsay-pub-grill/menus/clv-gr-pub-dinner-menu-071118.pdf

Le village buffet Official Site

https://www.caesars.com/paris-las-vegas/restaurants/le-village-buffet#.XA0BhafxJQI

Data resources -- Yelp dataset

https://www.yelp.com/dataset/challenge?fbclid=IwAR1_Y50tEjlemIMJxV68RkkdLCHRj1cQzl1DM9sBIFQ3Vo0JmgSPxIAaM5M

Topic Analysis: topic modeling

https://cfss.uchicago.edu/fall2016/text02.html#lda_with_an_unknown_topic_structure

Sentiment analysis

https://monkeylearn.com/sentiment-analysis/

Ggplot2

https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html#aesthetic_mapping

'tm' - Text Mining Package for R

https://cran.r-project.org/web/packages/tm/tm.pdf

as.Date function

https://www.statmethods.net/input/dates.html

Guide to the ngram Package

https://cran.r-project.org/web/packages/ngram/vignettes/ngram-guide.pdf

Package 'SnowballC'

https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf

Multiple plot in R

https://www.datamentor.io/r-programming/subplot/