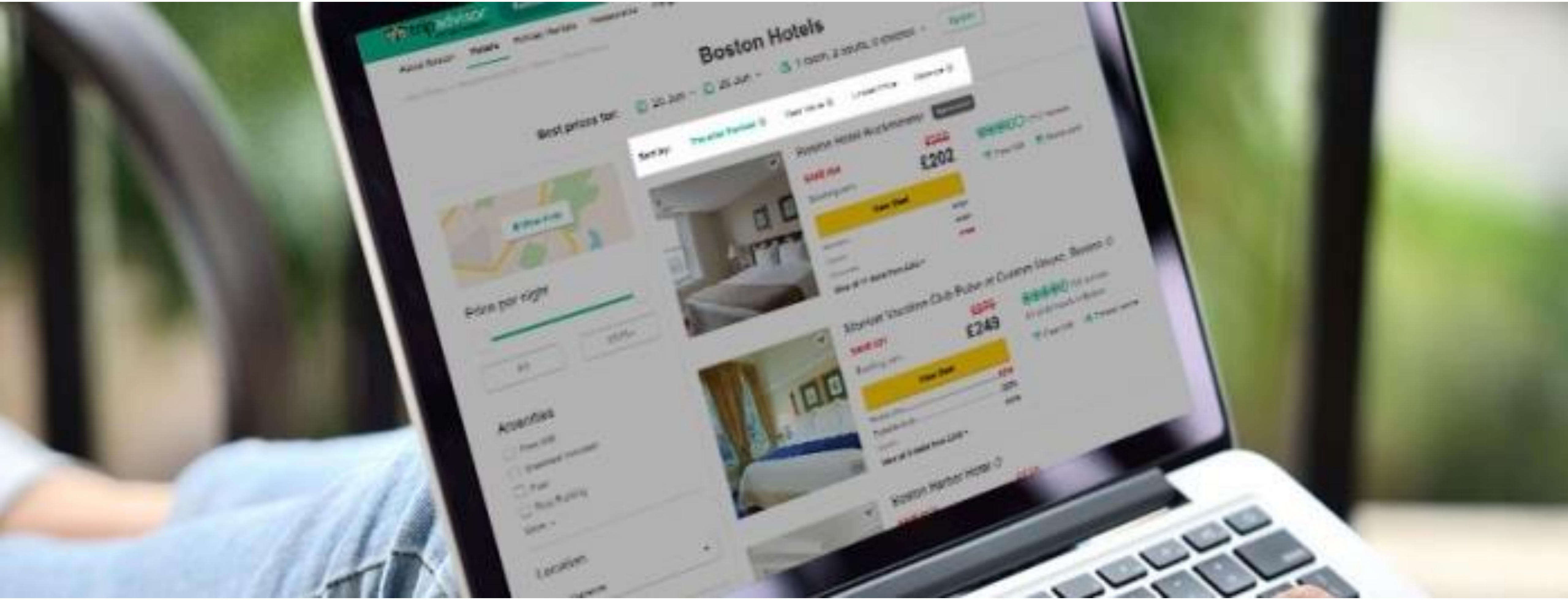
Group 5
Becky | Lawrence | Tony

Intelligence from reviews: the fundamental for 5 star restaurants

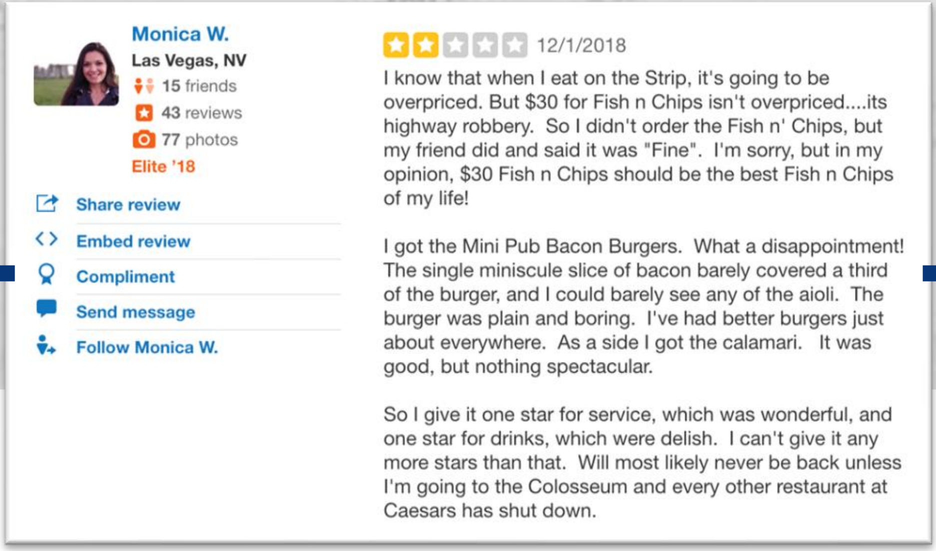
Introduction

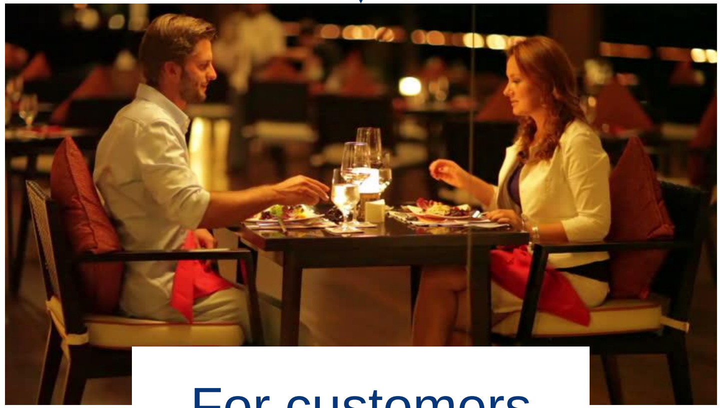
A problem that you and me will have



Introduction

Businesses have a larger incentive to analyse review data



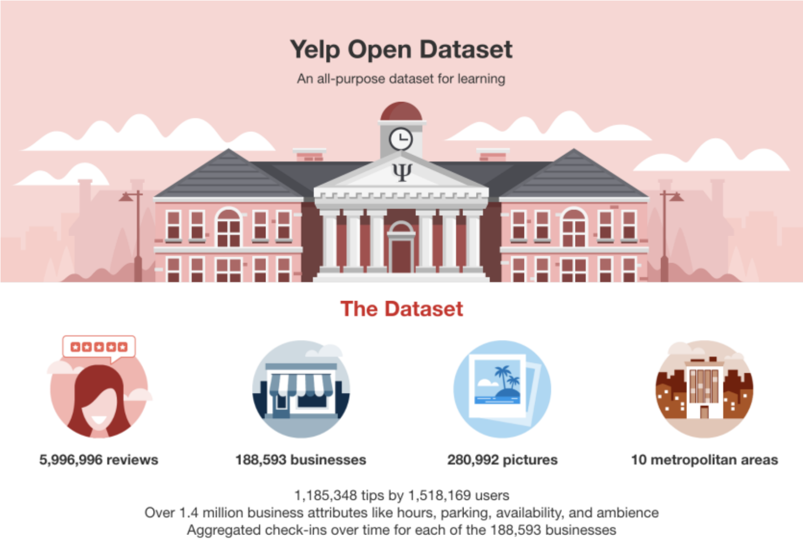


For customers

For businesses

Data Availability

Select restaurants from Yelp dataset with sufficient reviews



- ~70 restaurants with reviews > 2,000
- ~65 located in Las Vegas

Gordon Ramsay Pub & Grill

- English club restaurant
- 2,879 observations
- Average rating > 3

Le Village buffet

- French-styled buffet
- 2,246 observations
- Average rating =< 3

Our objective is to generate useful information for owners (e.g. Mr. Ramsay) to improve their restaurants

Hypotheses & Implications

We generated six initial hypothesis to guide our text analysis

1	Specific elements that most customers will be evaluating at	Advise the business focus on specific areas to enhance customer experience
2	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
3	Specific problems that will mentioned in majority negative reviews	Advise the business to solve the significant problems
4	Specific dishes that mentioned frequently in positive reviews.	Advise the business to put the photo upfront and advice customers what to order
5	High review frequency or good rating in specific holiday.	Advise the business to focus on particular period
6	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

Overview of Analysis

Three analyses covered in our project to identify meaningful data from reviews

problems, dishes, elements for good / bad reviews

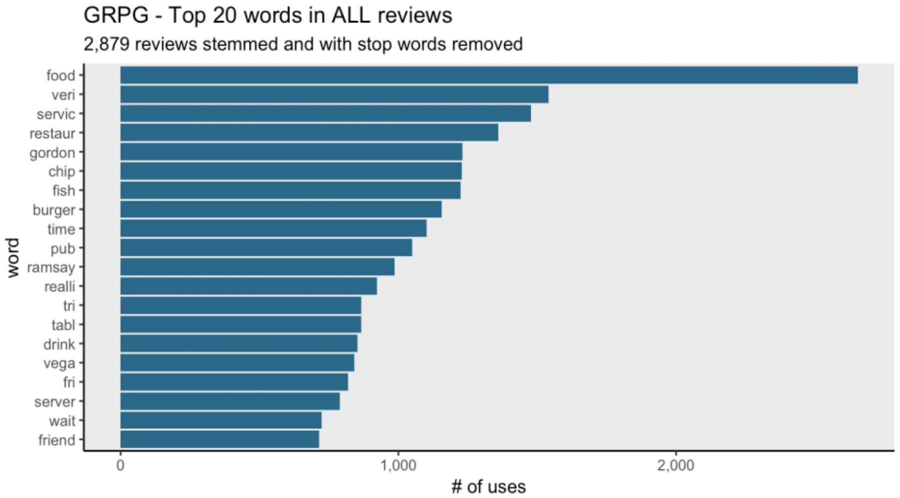
customer groups, feedback changes in special occasions

elements that customers evaluating on

1 – Opinion Mining

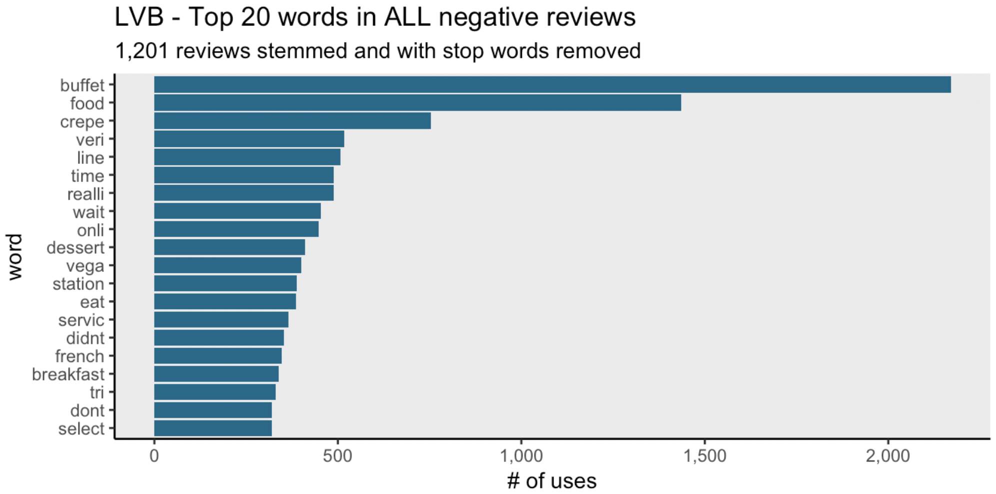
Classify generic, adverb and name related words to meaningless e.g. food, very, gordon

Find the most frequently mentioned words in different subset



Meaningful words:
“servic” (service), “chip”, “fish”, “burger”, “time”, “tabl” (table), “drink”, “fri” (fries), “server”

Meaningful words:
“crepe”, “line”, “time”, “wait”, “dessert”, “breakfast”

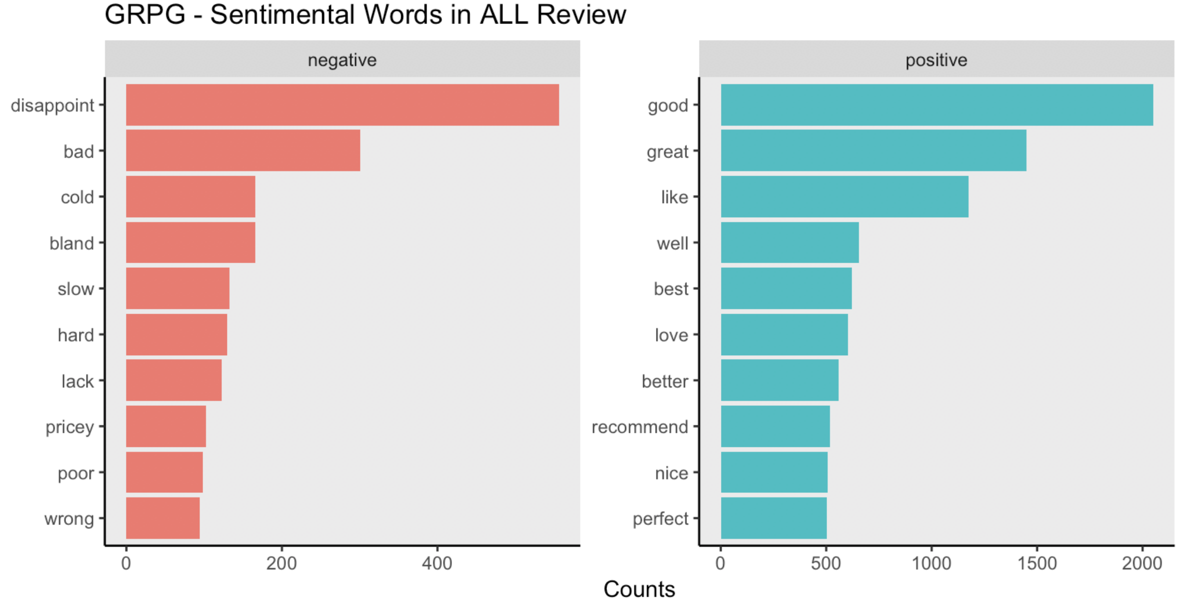


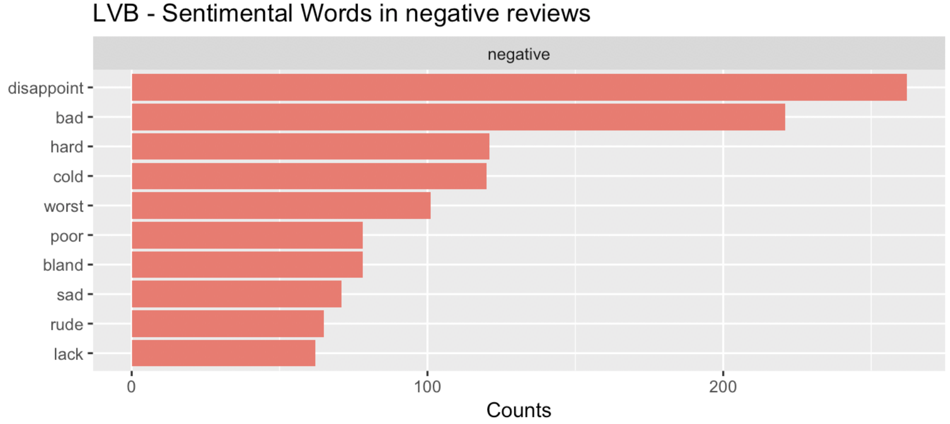
1 – Opinion Mining

Identify key sentimental words

Find the most frequently mentioned words in different subset

Summarize commonly used sentimental words that contribute to high and (or) negative reviews





1 – Opinion Mining

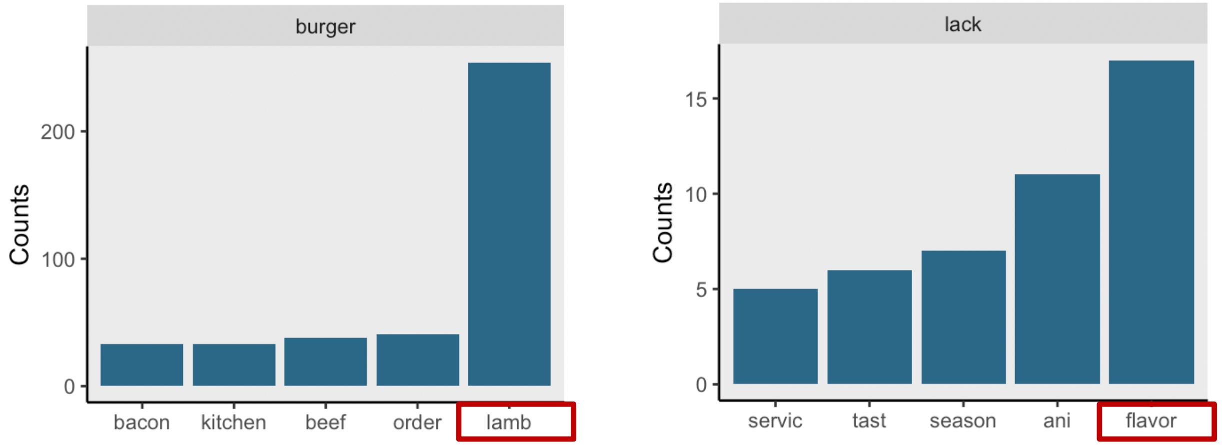
Bigrams to identify implications based on significant data

Find the most frequently mentioned words in different subset

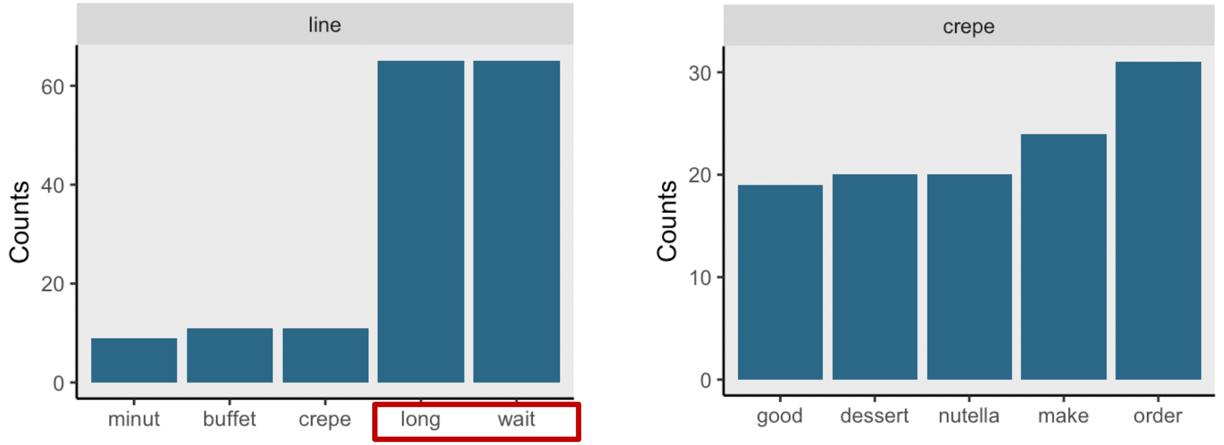
Summarize commonly used sentimental words that contribute to high and (or) negative reviews

Extract bigrams from the reviews and filter based on the objects or the sentimental words

GRPG



LVB



2 – Time-period Analysis

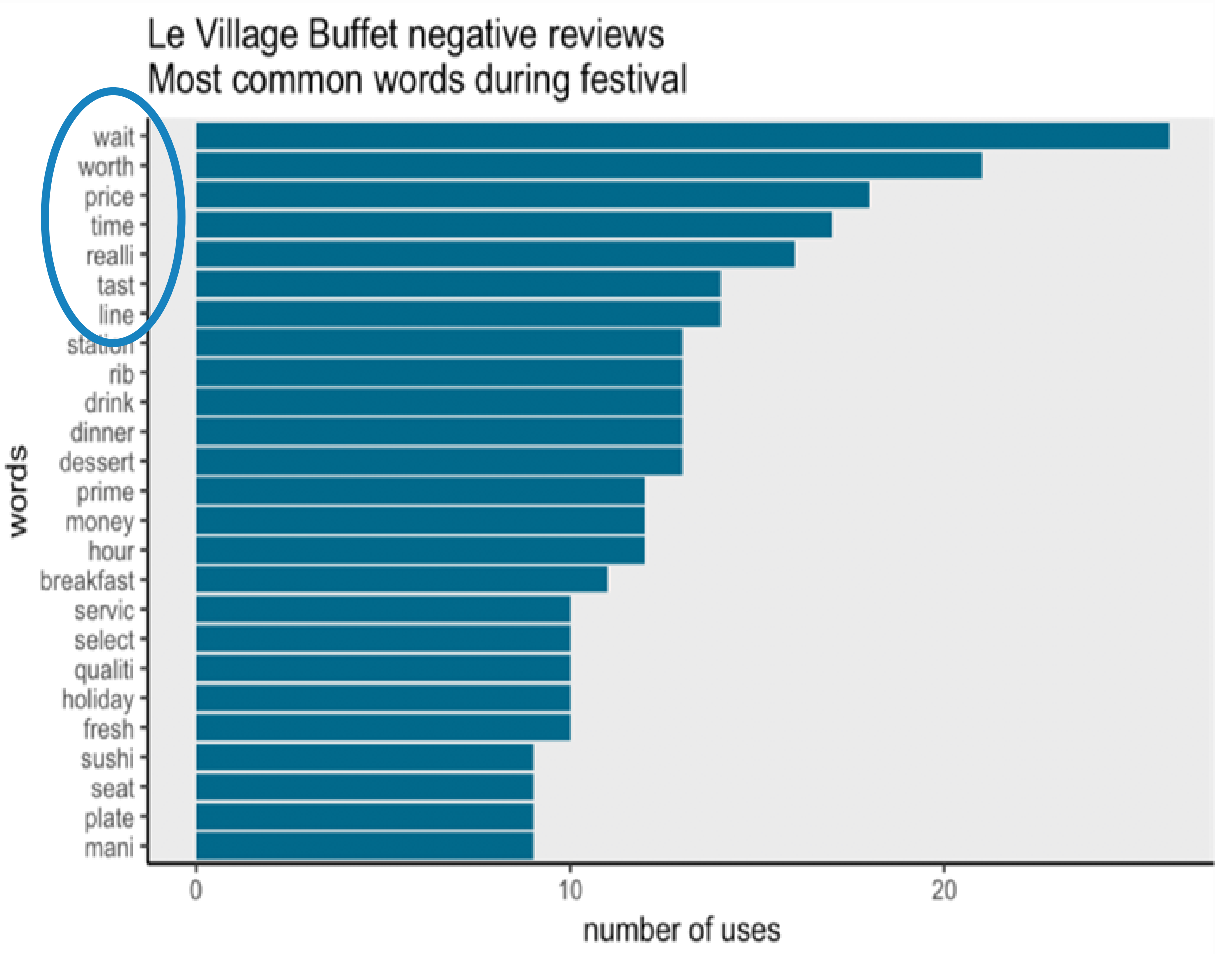
Do people rate higher during festival period?

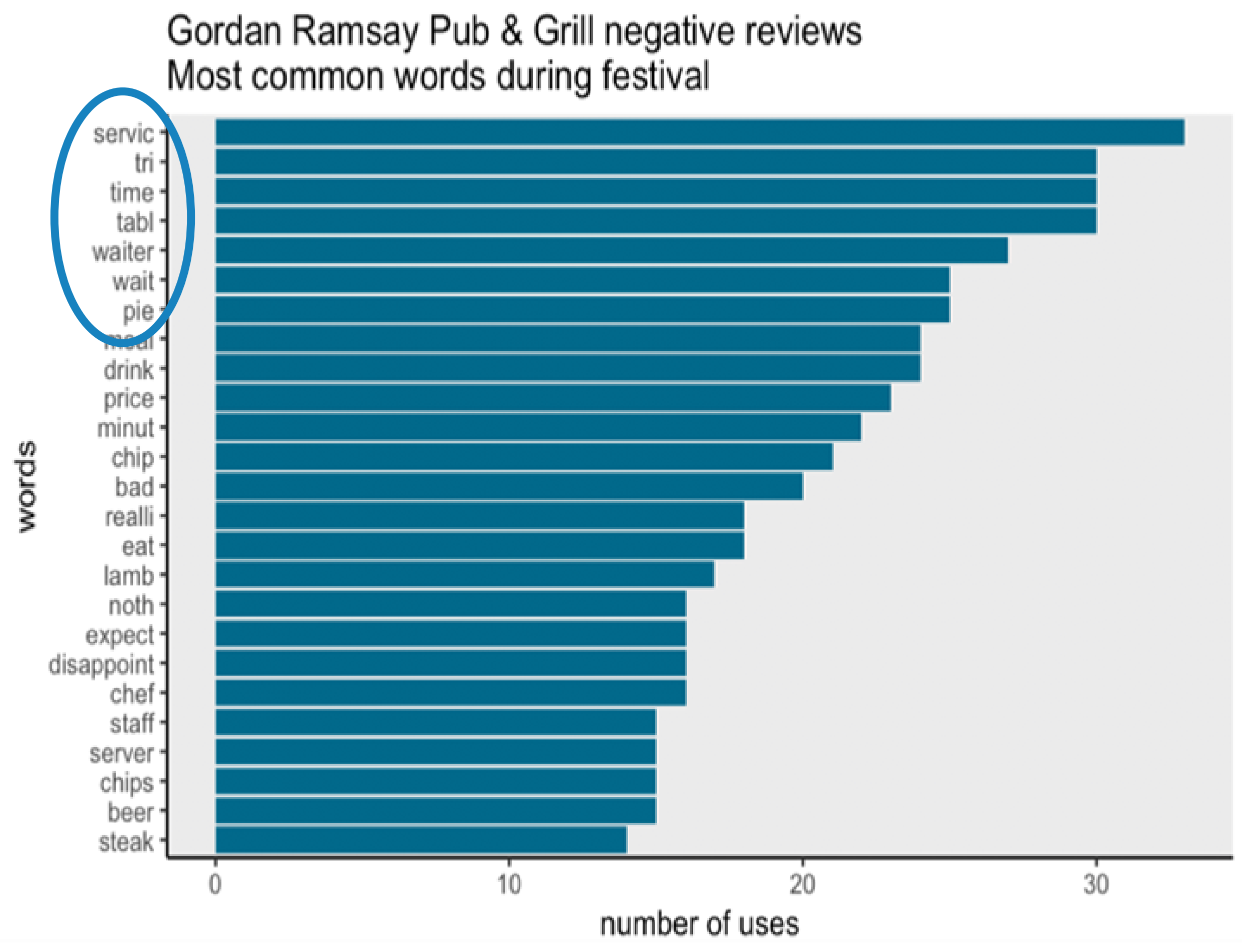
Original data → Festival Data → Compare the average rating

Rate of negative review (stars <= 3)	Festival Data	Original Data
GRPG	0.56	0.47
LVB	0.60	0.53

2 – Time-period Analysis

Why?? -- Words that are frequently mentioned in negative reviews during festival period.





3 – Classification

Unsupervised Machine Learning to classify review content

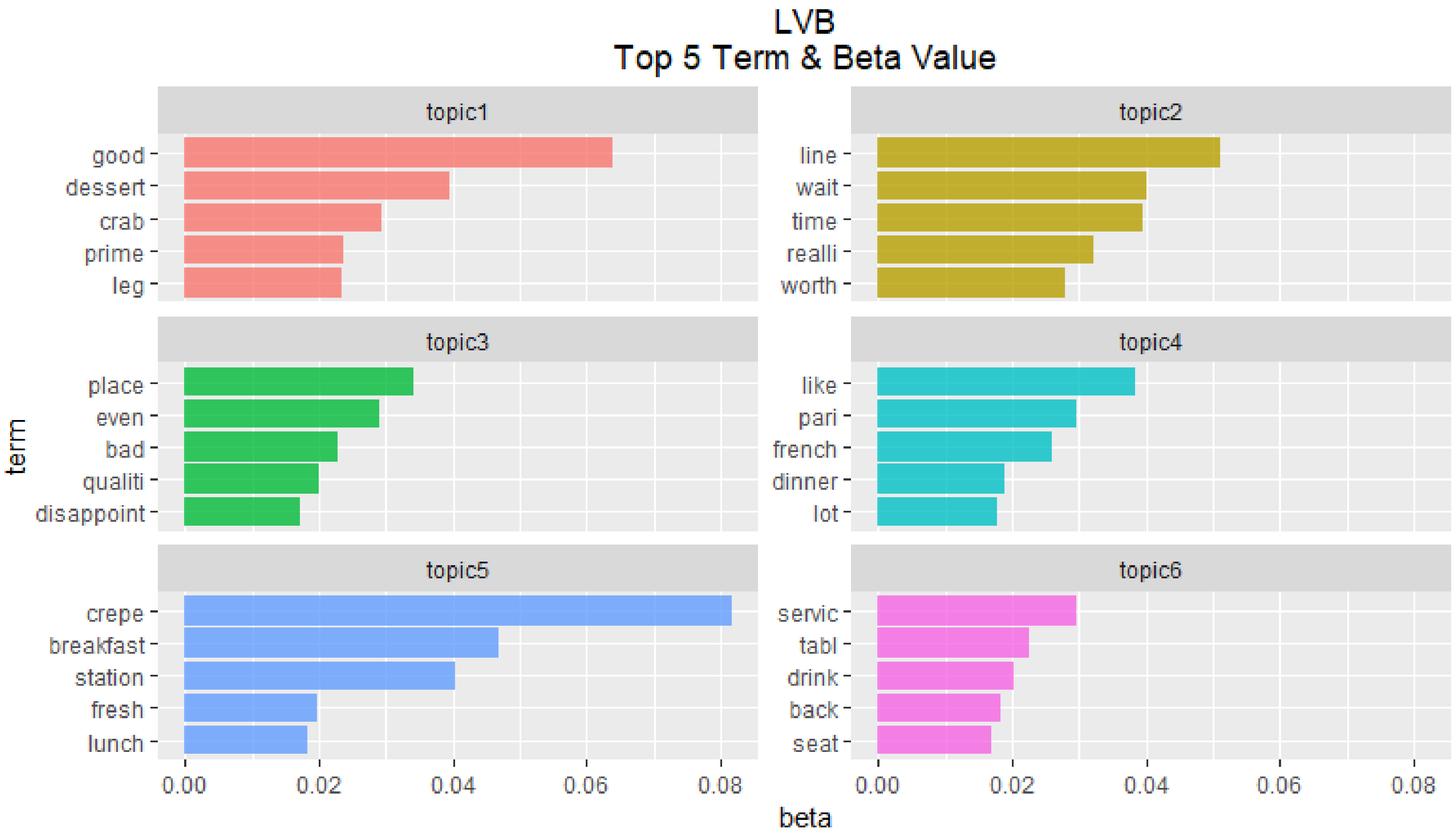
Data Pre-processing

Clean data and exclude stopwords
Create corpus
Generate Document-term Matrix (DTM

Decide Topic Number

Fit LDA model
Calculate beta value
Observe top terms
Try different numbers of topics
Select the optimal number

Analyze results



3 – Classification

Unsupervised Machine Learning to classify review content

Define Topics

Example.
“Table, Order, Wait, Ask, Waiter” in Topic 6
→ About “Taking orders”

Sentiment Analysis

Calculate posterior probability
Calculate sentiment score of each review
Removing outliers

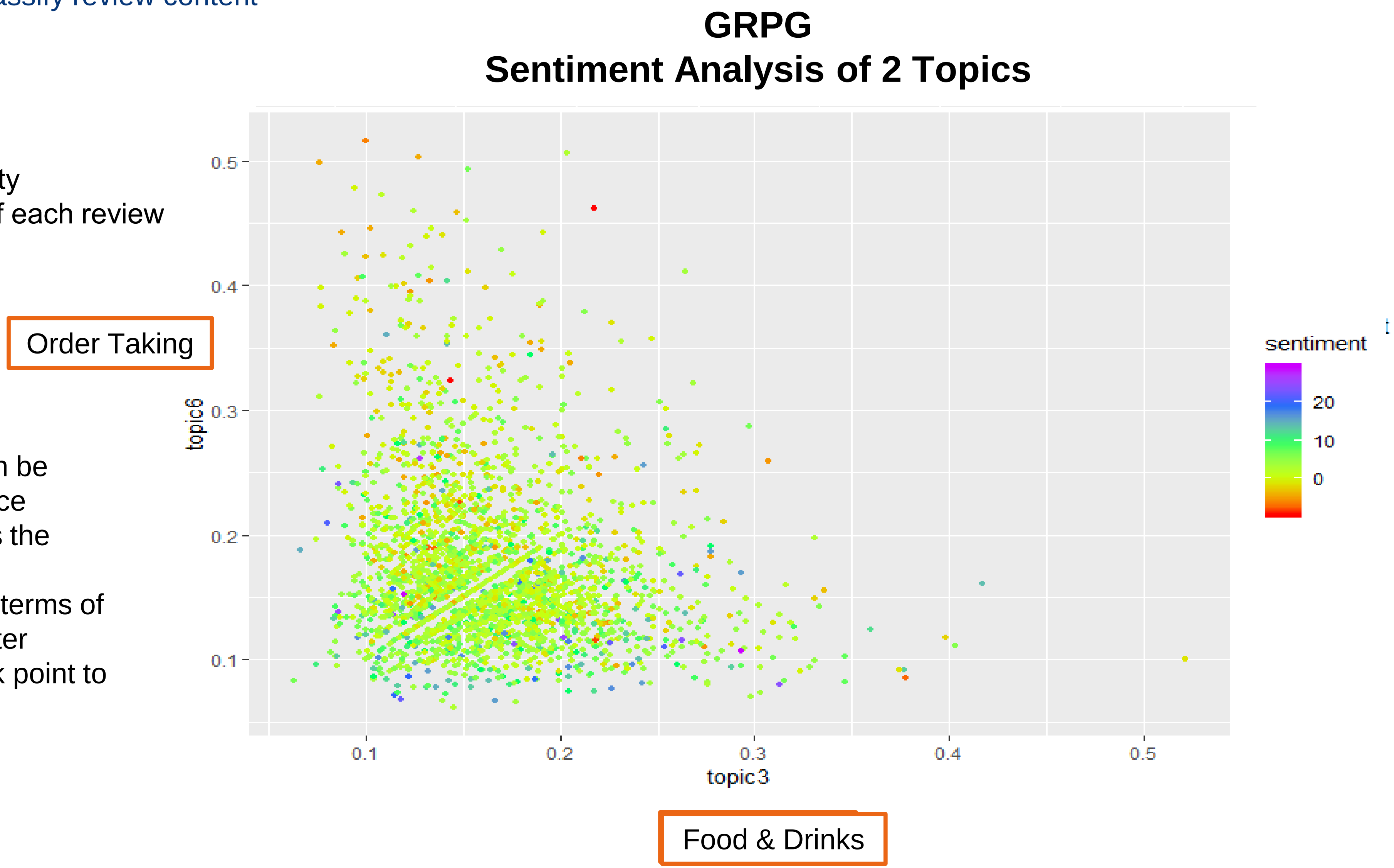
Plot the results

3 – Classification

Unsupervised Machine Learning to classify review content

Sentiment Analysis

Calculate posterior probability
Calculate sentiment score of each review
Removing outliers



Conclusion

- Order taking process can be improved while the service provided overall satisfies the customers.
- Order taking process, in terms of the waiting time and waiter performance, are a weak point to GRPG.

3 – Classification

Unsupervised Machine Learning to classify review content

Conclusion

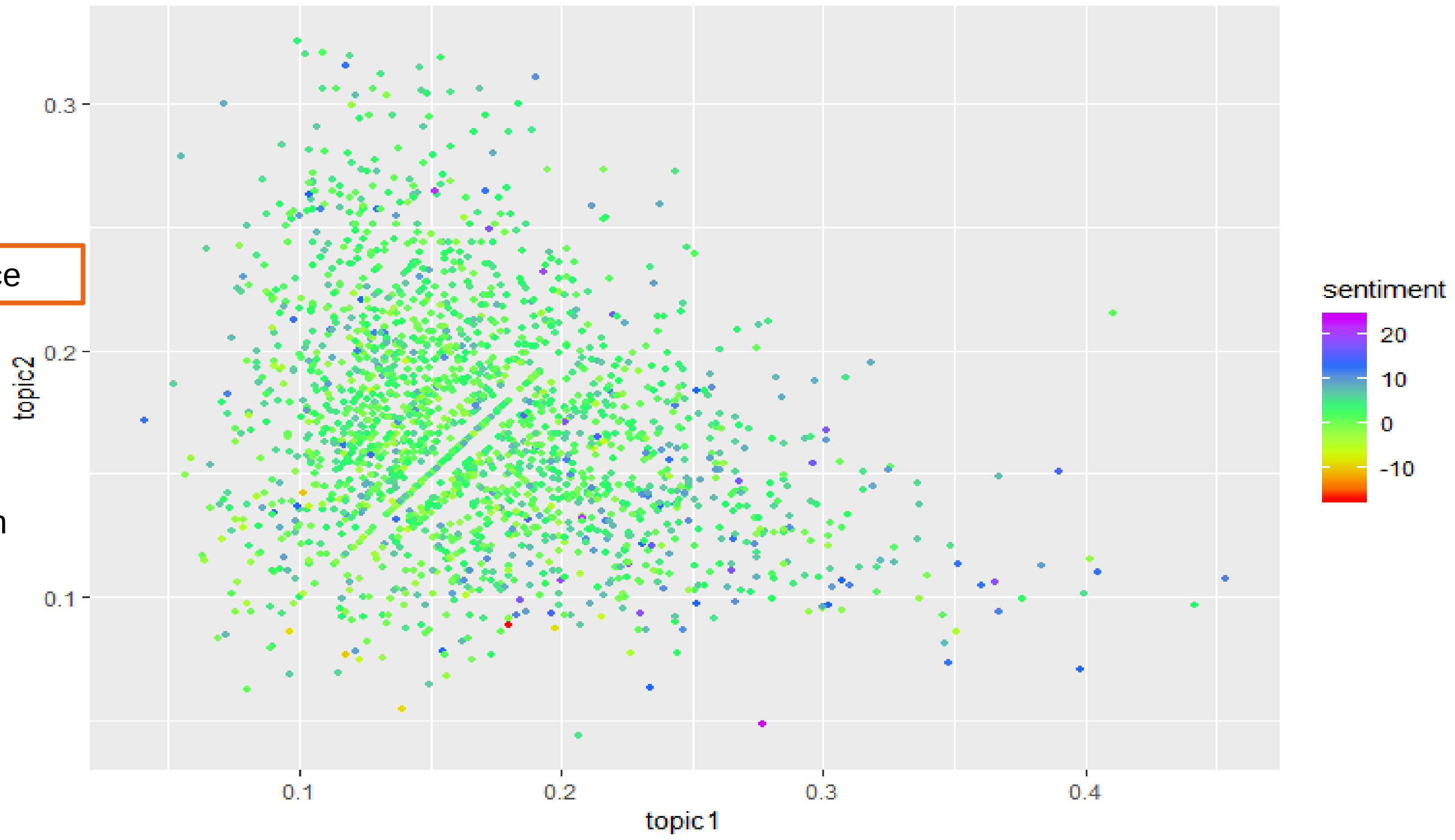
- there is no obvious weak points for LVB among food, serving and waiting time.

Limitations

- Contents highly similar
- Topic definition is subjective
- Not every 2 topics have relation with each other

LVB
Sentiment Analysis of 2 Topics

Service



Food & Desserts

Conclusion

Key findings

Opinion Mining		Time-period Analysis	Classification
Specific problems mentioned in majority negative reviews	Specific dishes mentioned frequently in positive reviews.	Elements that most customers will be evaluating at	Good rating in specific holiday
Lamb burger signature dishes for GRPG	LVB to focus on queuing in front of food stations	6 important topics for restaurants	bad rating are more frequent

Group 5
Becky | Lawrence | Tony

Intelligence from reviews: the fundamental for 5 star restaurants

Appendix: Data collection and pre-processing manipulation

review_id	user_id	business_id	stars	date	text
1EwQzhFsHX1C4-Zxs4PwVQ	QNH72vmMZMdyiuaZKh1l8A	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

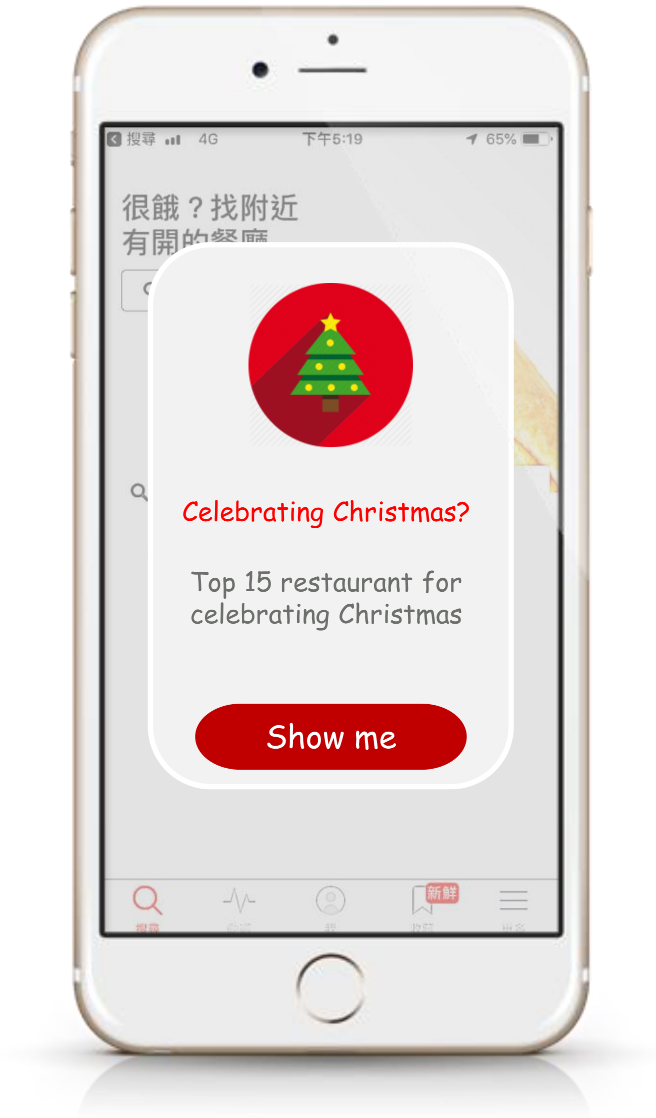
To prepare the review data for further analysis, our team applied the tm package to purify the text. Firstly, we used the tolower() function to convert all text into lowercase and str_replace_all() from stringr package to get rid of strange unicode in the text such as é, ñ and <U+00a0>. Then, we removed the punctuation using removePunctuation(), removed stopwords using removeWords() with the list from SnowballC package, stemmed the document using stemDocument().

Appendix: 2 – Time-period Analysis: Limitation

Appendix: 2 – Time-period Analysis: Limitation

What can we do if we reach this kind of data?

1. Compare review during festival period between restaurants
2. Find out the most popular restaurant during specific festival.
3. Pop out recommendation during festival period



Appendix: Data collection and pre-processing manipulation

Classification

```
setwd("D:/NTU/4_Exchange/HSG/Fall 2018/Quantitative Text Analysis/QTA_Final Project")

# Switch different restaurants

# grill
data <- read.csv("review_data_GR.csv")
data <- data[data$business_id == "YJ8ljUhLsz6CtT_2ORNFmg", ]

# buffet
#data <- read.csv("set3.csv")
#data <- data[data$business_id == "ZkGDCVKSdf8m76cnnalL-A", ]

library("dplyr")
library("tm")
library("readr")
library("stringr")
library("textstem")
library("corpus")

# Define stopwords for restaurant review analysis
defined_stopwords <- c("meal", "eat", 'burger', "food", "gordon", "ramsay", "ramsey","vegas","vega","one", "think",
                       "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")

defined_stopwords2 <- c("meal", "eat", "food","one", "think","village","vega","vegas","buffet","le",
                        "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")

```

Appendix: Data collection and pre-processing manipulation

Classification

```
# Data Pre-processing
clean_data <- data %>%
    mutate(text = as.character(text)) %>%
    mutate(text = removeNumbers(text)) %>%
    mutate(text = tolower(text)) %>%
    mutate(text = removePunctuation(text)) %>%
    mutate(text = stripWhitespace(text)) %>%
    #mutate(text = lapply(text, unique)) %>%
    mutate(text = lemmatize_strings(text)) %>%
    mutate(text = removeWords(text, stopwords("english"))) %>%
    mutate(text = removeWords(text, defined_stopwords)) %>%
    mutate(text = text_tokens(text, stemmer = "en")) %>%
    mutate(text = substring(gsub(",", "", gsub("\"", "", str_c(text))), 3))
#  mutate(text = str_replace_all(text, "\\s", " ")) %>%

#clean_data$review_id= NULL
clean_data$user_id = NULL
clean_data$X = NULL


# Create Document-term Matrix
#DTM_matrix <- strsplit(as.character(clean_data$text), "\\s+")
myCorpus <- Corpus(VectorSource(clean_data$text))
review_matrix_counts <- DocumentTermMatrix(myCorpus)
rowTotal <- apply(review_matrix_counts, 1, sum)
review_matrix_counts <- review_matrix_counts[rowTotal > 0,]


# Calculate term frequency
counts <- colSums(as.matrix(review_matrix_counts))
counts <- sort(counts, decreasing = TRUE)
```

Appendix: Data collection and pre-processing manipulation

Classification

```
library(topicmodels)

# fit LDA model
review_LDA <- LDA(review_matrix_counts,
                 method = "Gibbs",
                 k = 6,                     # suppose we have 5 topics
                 control = list(seed = 1234))

#terms(review_LDA)
#topics(review_LDA)

library(tidytext)
betaMatrix <- tidy(review_LDA, matrix="beta")

topTerms <- betaMatrix %>% group_by(topic) %>% top_n(15) %>% ungroup() %>% arrange(topic, -beta)
topTerms

library(tidyr)
beta_spread <- betaMatrix %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic2 > .001 | topic3 > .001 | topic4 > .001 | topic5 > .001| topic6 > .001 )#| topic7 > .001 | topic8 > .001)

# Selecting best topic setting for word
beta_spread$bestTopic = names(beta_spread)[apply(beta_spread, 1, which.max)]
beta_spread = mutate(beta_spread, beta = (pmax(topic1,topic2,topic3,topic4,topic5,topic6)))#,topic7,topic8, topic9)))#, topic10)))#) / s
# Removing redudant topic columns ("2 = topic1" ~ "7 = topic6")
beta_spread <- beta_spread[, -c(2:7)]

# Group by best topic fit and selecting the top 5 words
beta_spread <- beta_spread %>% group_by(bestTopic) %>% top_n(5)
```

Appendix: Data collection and pre-processing manipulation

Classification

```
# Plot term-beta graph of 8 topics
library(ggplot2)
beta_spread %>%
    mutate(term = reorder(term, beta)) %>%
    ggplot(aes(reorder(term, beta), beta, fill = factor(bestTopic))) +
    ggtitle("LVB\nTop 5 Term & Beta Value") +
    theme(plot.title = element_text(hjust = 0.5)) +
    geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
    facet_wrap(~bestTopic, scales = "free_y", ncol = 2) +
    coord_flip() + xlab("term")

review_document = tidy(review_LDA, matrix = "gamma")


# Calculating the topic probablity for each review
topics <- posterior(review_LDA)$topics
colnames(topics) <- paste("topic", 1:6, sep = "")

sentiment_data <- clean_data %>%
    unnest_tokens(word, text) %>%
    inner_join(get_sentiments("bing")) %>%
    count(review_id, sentiment) %>%
    spread(sentiment, n, fill = 0) %>%
    mutate(sentiment = positive - negative)

# Combining original data, topic probability, and sentiment data
combined <- merge(cbind(clean_data, topics), sentiment_data, by = "review_id") %>%
  filter(sentiment < 30 & sentiment > -20) # Remove outliers

# Ploting out the data with the probability of the two chosen topics as the axis, and the sentiment as the color scale
ggplot(combined, mapping = aes(x = topic3, y = topic6, color = sentiment)) + geom_point(size = 1) +
    scale_color_gradientn(colours = rainbow(5))
```

Appendix: Data collection and pre-processing manipulation

Classification

Beta Spread

	term	bestTopic	beta
1	amaz	topic1	0.02632795
2	ask	topic6	0.02058872
3	beer	topic3	0.02435069
4	chip	topic5	0.04865329
5	dessert	topic4	0.01772203
6	disappoint	topic2	0.02297142
7	expect	topic2	0.02675859
8	fish	topic5	0.04791441
9	fri	topic5	0.03239794
10	friend	topic1	0.02983504
11	good	topic2	0.08649813
12	great	topic1	0.06111006
13	grill	topic3	0.01687108
14	littl	topic5	0.02193695
15	menu	topic3	0.02706013
16	order	topic6	0.02761032
17	perfect	topic4	0.01958510
18	pie	topic4	0.02620092
19	place	topic2	0.04487999
20	price	topic2	0.02590342
21	pub	topic3	0.04056921

Review document

	document	topic	gamma
1	1	1	0.22456140
2	2	1	0.22701149
3	3	1	0.15873016
4	4	1	0.20531401
5	5	1	0.24731183
6	6	1	0.14859438
7	7	1	0.10397554
8	8	1	0.22695035
9	9	1	0.13468013
10	10	1	0.08888889
11	11	1	0.16246499
12	12	1	0.19047619
13	13	1	0.14784946
14	14	1	0.16838488
15	15	1	0.29535865
16	16	1	0.24472574
17	17	1	0.13963964
18	18	1	0.12009804
19	19	1	0.18473896
20	20	1	0.19540230
21	21	1	0.20451527

Appendix: Hypotheses checklist

We generated six initial hypothesis to guide our text analysis

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