Restaurant Online Reviews and Their Manipulations

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Applied Economics

in the

College of Graduate Studies

University of Idaho

by

Guang Tian

Major Professor: Liang Lu, Ph.D.

Committee Members: Christopher McIntosh, Ph.D.; Jason Winfree, Ph.D.

Department Administrator: Christopher McIntosh, Ph.D.

August 2020

Authorization to Submit Thesis

This thesis of Guang Tian, submitted for the degree of Master of Science with a Major in Applied Economics and titled "Restaurant Online Reviews and Their Manipulations," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor:	Liang Lu, Ph.D.	7/24/2020 Date
Committee Members:	Christopher McIntosh, Ph.D.	7/27/2020 Date
	Jason Winfree Jason Winfree, Ph.D.	7/27/2020 Date
Department Chair:	Christopher McIntosh, Ph.D.	7/27/2020 Date

Abstract

My thesis contains two separate chapters. In the first chapter, I focus on what consumers express about their dining experience. I present a case study using Yelp restaurant review data on what affects restaurant customers' sentiment responses to dining out. The case study uses a sentiment-analysis method to extract and generate sentiment indices from 175,879 text-based Yelp restaurant reviews. I then examine the relationships between the sentiment indices and ratings and what factors(food, service, expenditure, and social) affect consumer dining experience more than others. In the second chapter, I find that online consumer reviews can be dependent upon the type of review platform. Since online review platforms vary, platforms will have different subsets of customers leave reviews and the detail of those reviews may vary. Platforms that attract more satisfied consumers will tend to have more inflated averages. Previous research has shown that platforms may have fake reviews, but given platform differentiation, even if all the reviews are real, consumers can still receive biased information from online reviews. These findings are beneficial for online platform design and provide empirical evidence for information design theory.

Acknowledgements

First of all, I would like to sincerely thank my major professor, Dr. Liang Lu, for his dedicated support and guidance. I was not only impressed by his comprehensive knowledge but also his work enthusiasm. He is always more than happy to encourage me to explore my interesting research topics and providing me guidance without any reservation. It is his encouragement that makes me more decisive about choosing academia position as my future career.

Second, I would like to express my gratitude to my two other committee members, Dr. McIntosh and Dr. Winfree. Dr. McIntosh provides me consistent supports throughout my master's studying period. Especially, I would like to thank him for supporting me to present my papers at AAEA and WAEA annual meetings. Without his help, I would lose these important opportunities to improve the my papers and communicate with other scholars. Also, I would like to express my appreciation to Dr. Winfree who gives me critical help for the theoretical section of my thesis. I admire his sharp insights and model-constructing skills.

Last but not the least, I must thank my parents and family for their consistent support for my life and study. During this special COVID-19 period, I miss them more than ever. This thesis is dedicated to my beloved parents. I hope everyone is staying healthy.

Table of Contents

List of Tables

List of Figures

Chapter 1: What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data

1.1 Introduction

A research interest focused on the use of big data in sensory and consumer science has been steadily growing in the recent years. For example, researchers have started to look at new data sources (e.g., social media, Vidal et al., 2015) as alternative or complementary to traditional consumer research. They are also looking at new types of variables (e.g., emoticons and text from tweets, Vidal et al., 2016; Schouteten et al., 2019) to measure emotions. In addition, new research tools (e.g., machine learning Heng et al., 2018) are being used to analyze these new types of data.

The goal of this paper was to measure food-related consumer sentiments and affective responses from online review data. Online reviews refer to customer opinions and experiences posted on online review or retailing platforms. For example, Yelp enables its users to leave ratings and text reviews about their dining experiences at restaurants. Grocery shoppers on Amazon or Walmart can give ratings and reviews about the food or beverages they purchased. In this paper, I used Yelp restaurant review data to analyze food-related sentiments. From the new data, I focused on consumers' text reviews where sentiments were more explicitly documented than the overall star-rating score. I used sentiment analysis and semi-automatic content analysis as tools to measure and analyze sentiment words expressed in the text reviews.

There are three advantages of using online review data to study consumer dining sentiments. First, it provides spontaneous consumer sentiments in natural consumption settings. This can provide more realistic data about consumer dining experience compared with other kinds of data. Benefits of using spontanous consumer data has been firmed by Vidal et al. (2015). Second, it contains comprehensive information. For example, in the Yelp dataset that I used, contained consumer characteristics like review count, user account age, etc. The Yelp data also contains business features like parking, allows pets, has WiFi availability, etc. Third, it is supported by large sample sizes, and are readily available on many major platforms. For example, in the Yelp dataset there are millions of restaurant related reviews. Apart from Yelp, there are many other platforms that can provide restaurant review information, like Opentable, Google My Business, etc.

Previous research using online review data can be broadly divided into two categories: (i) computer science research focusing on developing big data algorithms that used online review data as tools to test the algorithms (see, e.g., Carbon et al., 2014; Huang et al., 2014; Yu et al., 2017) and (ii) business and marketing related research to predict consumer choice, product sales and make marketing strategy decisions (see Chevalier and Mayzlin, 2006; Liu, 2006; Anderson and Magruder, 2012; Liu and Park, 2015; Luca, 2016; Felbermayr and Nanopoulos, 2016; Campbell et al., 2017). However, literature using online review data to analyze food-related consumer sentiment is relatively scarce. To the best of my knowledge, the research that most closely match the current study was done by Nakayama and Wan (2018, 2019). Nakayama and Wan (2018) studied the different sentiment words usage between American and Japanese consumers toward five dining factors (food quality, service, ambience, price fairness, and others) of Japanese restaurant. In Nakayama and Wan (2019), they further considered the different usage of sentiment words among 10 kinds of Japanese cuisine (Ramen, Sushi, Tempura, etc.). There are two main differences between my paper and Nakayama and Wan (2018, 2019). First, rather than using consumer rating as a standard to identify positive and negative sentiment (4 and 5 star were considered as positive sentiment; 1 and 2 star were considered negative sentiment), I extracted and generated sentiment indices from text reviews using sentiment analysis. Second, I used a restaurant fixed effect model to consider the relationship between sentiment words in a customer's review and rating, and to evaluate/assess what factors can affect the usage those sentiment words.

My contributions to sensory and consumer science literature were three-fold: First, I provided insight into consumer sentiment studies from a new angle by using online review data. Second, I identified the topics (food, service, social, and expenditure) which were linked to consumers' usage of sentiment words. Third, I further identified the relations between these topics and consumers' rating of restaurants.

1.2 Materials and Methods

Data

The Yelp Dataset Challenge program provides access to the Yelp review dataset in 10 metropolitan areas in the United States and Canada.[1] The dataset is updated every year and contains 4.5 million reviews (see Figure B.1 for an example) across local businesses such as restaurants, shopping malls, and services, all of which were generated by Yelp users from January 2005 to December 2017.

The master data file contained six subsets. "Yelp review" contained variables such as business ID, stars (a business's numerical rating), date, and text review. "Yelp business" contained variables such as business ID, name, address, city, postal code, and categories. "Yelp business attributes" contained business ID and other 81 variables describing features such as whether the business has parking, allows

[1] https://www.yelp.com/dataset/challenge. Last accessed 09/18/2019.

pets, has WiFi availability, etc. "Yelp check-in" shared users' locations on their social media account. "Yelp tip" was only available through the Yelp mobile application and displays high-frequency words extracted from the text reviews. "Yelp user" contained variables such as review count, user account age, friends, and elite status (Yelp certified, high-quality review provider), and average stars by the user.

Figure 1.1: Yelp customer review example

Note: User name, user location, user photo, and restaurant name were intentionally removed.

I used review data from the Yelp restaurant dataset for the city of Las Vegas, Nevada. I used the data from Las Vegas rather than from all 10 cities in the dataset because, first—and unlike most other cities included in the dataset–Las Vegas is primarily a tourist destination. Such consumers are less likely to reside in the city and therefore rely more on online review platforms rather than their past experiences or the experiences of other locals. Local consumers, who have past experience or established expectations for certain restaurants, are less likely to post reviews when a dining experience is close to their expectations. Consequently, reviews from local consumers tend to be more extreme (either positive or negative) and become less representative. Second, processing all 10 cities' worth of data (4.5 million observations) would require higher computational capacity than I had available.

I took four steps to construct my dataset from the original Yelp dataset: First, I selected all the restaurants in Las Vegas from 2005 to 2017 contained in the "Yelp business" subdataset and obtained a list of all the restaurant names and their unique business IDs assigned by Yelp. Second, using the unique business IDs, I merged the "Yelp business" and "Yelp review" subdatasets to obtain a panel dataset containing 310 Las Vegas restaurants with 175,879 reviews from 2005 to 2017. The unit of observation was review by day. Third, I used textual analysis tools (discussed in section 2.3) to generate sentiment indices and topic variables from the text reviews. Finally, I collapsed the daily observations into monthly averages to deal with the fact that many restaurants were not reviewed on a daily basis. After data processing, the dataset contained 20,864 observations from 310 Las Vegas restaurants over 156 months (from January 2005 to December 2017). The panel dataset was unbalanced due to restaurants entering and exiting the market.

Data processing

The dataset was cleaned and maintained by Yelp.com. The text reviews were processed by the Linguistic Inquiry and Word Count (LIWC) software (introduced in section 2.3), which automatically performs tokenization, stemming, and stop words removal. I did not perform a normalization step on the text. Instead, I normalized the numerical sentiment indices generated from the sentiment analysis.

Sentiment analysis

In order to extract positive and negative sentiment words contained in a restaurant review text, I used a lexicon-based sentiment-analysis method (Pennebaker and Francis, 1996; Tausczik and Pennebaker, 2010; Pennebaker et al., 2015), which involved two important steps: First, I used a dictionary, constructed by linguistic researchers, of positive and negative sentiment words. Second, I calculated the number of positive and negative sentiment words used in a review to generate positive and negative sentiment indices for the text review. I employed the LIWC text analysis software dictionary—which draws from up-to-date linguistics and psychology research—to identify the sentiment words from each restaurant review text. The dictionary contained basic emotion and cognitive words often used in social, health, and personality psychology studies as well as words commonly used in text messages and social media posts. The dictionary contained 6,400 words (including word stems),[2] which had been categorized into a wide range of psychological groups (e.g., affection, cognition, logic). The word stems

[2]For example, enjoy* is a word stem. The LIWC text analysis software will match any words with the same first five letters (e.g., enjoyable, enjoyment)

were used to compute the sentiment indices.

For my research, I were interested in categories of positive sentiment words and negative sentiment words, which contained 406 words (e.g., :), happy, lmfao) and 499 words (e.g., :(, piss, lame), respectively. It is worth noting that not only words that could be linked to emotions (e.g., happy, sad) were contained in the dictionary but also words that express positive or negative affective responses to the restaurant experiences (e.g., delicious, pricey). These affective response words in the LIWC dictionary were developed through sentiment analysis related linguistic research (e.g., Pennebaker et al., 2015). Sentiment analysis allowed one to capture consumer's dining sentiments through counting direct emotion words, and through affective response words which could indirectly reflect a consumer's dining experience. The validity of the dictionary had been cross-checked by experts by using different types of text materials, including social media posts, news, and historical documents.

As defined by LIWC, a positive (negative) sentiment index was the fraction of positive (negative) words in the total number of words used in a of review. For example, in the Yelp review shown in figure B.1,

> *Everything was delicious!! The staff has wonderful service. Would come back here everytime I'm at the [restaurant name][3]!! Love the food.*

"delicious," "wonderful," and "love" were counted as positive sentiment words, and there were no negative sentiment words. The review had 21 words, so the positive sentiment index was 14.29 ($(3/21) \times 100$) and the negative sentiment index was 0 ($(0/21) \times 100$).

Semi-automatic content analysis

By definition, "content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2018, p. 24). Content analysis has been used before to study food-related consumer behaviour. For example, Vidal et al. (2015) used content analysis to manually categorize tweets into six themes and 15 subthemes and then analyzed what people express in four dining contexts (breakfast, lunch, dinner, and snack). However, content analysis is a time-consuming process. If Vidal et al. wanted to analyze their whole samples, the categorization process would take two researchers 840 hours each to manually identify the themes of 64,000 tweets. In the end, their strategy was to randomly choose 16,000 of 64,000 tweets to analyze. Analyzing tweets is already time consuming, but using content analysis on online reviews data will take researchers even more time, for two reasons: (1) Text reviews have a higher average word count (129 words on average for Yelp reviews) than tweets (a maximum of 280 characters during the

[3]Restaurant name was intentionally removed. One potential concern was that whether a restaurant's name could contain sentiment words. In that case, a consumer's reference to the restaurant's name can be erroneously counted in the sentiment indices. In fact, less than 1 percent of restaurant name contains sentiment words.

period under analysis). (2) My study includes 175,879 text reviews, almost twice as many as in the study by Vidal et al.

To address the time-consuming nature of content analysis, I used a compromise content analysis method, a semi-automatic content analysis method. Wu et al. used this method to study gender discrimination in posts from an anonymous online forum. Instead of manually categorizing all of the posts, Wu et al. chose the 10,000 most frequently used words in all posts and manually categorized them into 15 themes with the help of canned software. Following Wu et al., I categorized the most frequently appearing words into predefined groups and identify words in each review that fall into these categories.

For this article, the categorizing work was done by one researcher with more than two years experience with food-related research and checked by two other researchers. First, I counted the word frequency of the all 175,879 text reviews and kept the top 300 with the most frequently used words (see Appendix A for a list of word frequencies).[4] Figure 1.2 displays a word cloud of the top 300 words. Second, I excluded sentiment words from the word list. After excluding sentiment words, I got 263 high-frequency topic words, which I categorized into five groups: "food," "service," "expenditure," "social," and "miscellany." Table 1.1 reports samples of each category (see Appendix B for details). For each text review, I counted occurrences of topic words in each category[5]. Then, for each of the five groups, I defined the value of group words as the fraction of high-frequency topic words in that category to the total high-frequency word count across all the categories. The sum of the values of the five group variables should be 1. For example, there were three high frequency topic words in the sample review text from section 2.3: "staff" and "service" (in the "service" category) and "food" (in the "food" category). Hence, the value of "food" was 1/3, the value of "service" was 2/3, and the values of "expenditure," "social," and "miscellany" were 0/3.

Table 1.1: Example words of each topic category.

Category	No. Words	Examples
Food	84	chicken, dinner, steak
Service	17	server, waiter, staff
Expenditure	5	worth, expensive, money
Social	8	friend, people, husband
Miscellany	149	give, look, star

Notes. I first counted the most frequent 300 words of all text reviews. Then I excluded sentiment words from the frequency word list. After excluding sentiment words, we got 263 high-frequency topic words, which I categorized into the five groups. See Appendix B for details about words in each topic category.

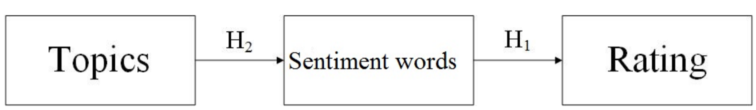
[4] "Stop words" such as "in," "a," and "the" were not included in the word-frequency count.
[5] The 263 high-frequency topic words covered more than 99% reviews in my dataset.

Figure 1.2: The most frequent 300 words in Yelp text reviews

Hypotheses

The purpose of this paper was to provide insight into consumer sentiment studies from a new angle by using the sentiment analysis method. As shown in Figure 1.3, my two specific tasks are (1) show that consumer sentiments in text reviews are linked to rating. (2) review topics (food, service, social, and expenditure) are related to consumer sentiments.

Figure 1.3: Hypothesis testing



First, by using survey data, previous scholars had shown that positive sentiments were positively correlated with food rating and negative sentiments were negatively correlated with food rating (Edwards et al., 2013). I wanted to show that the usage of sentiment words extracted from online restaurant text reviews also had such relations with numerical rating. To be specific, I wanted to show that higher ratings are related to a higher proportion positive sentiment word usage in text reviews and a lower proportion of negative sentiment word usage.

Hypothesis 1 *H_1: A rating is positively (negatively) related to the positive (negative) sentiment index.*

I tested Hypothesis 1 using a fixed-effects model, a widely used econometric technique for analyzing panel data (see, e.g., Piepho and Kalka, 2003). The advantage of using a fixed-effects model is that it will mitigate bias and endogeneity issues caused by unobservable restaurant characteristics. For example, factors like restaurant location, decoration, or cuisine style (e.g., French, Italian, Japanese) could affect both a consumer's sentiment and rating for the restaurant. By using a fixed-effects model, I can control for such factors even though these variables are not included in the dataset. My estimating equation is

$$Rating_{it} = \beta_0 + \beta_1 SPSI_{it} + \beta_2 SNSI_{it} + \alpha_i + \epsilon_{it}, \tag{1.1}$$

where $Rating_{it}$ refereed to the rating of restaurant i at time t; $SPSI$ and $SNSI$ were the standardized positive and negative sentiment indices, respectively;[6] α_i was the restaurant fixed effect for each restaurant i; and ϵ_{it} was the disturbance term.

Second, one of the goals of this paper was to inform researchers about factors that affected consumer sentiment word usage in online reviews. Previous research had already shown that understanding consumer sentiment was important for business success because consumer purchasing decisions were affected by emotions (Ladhari et al., 2008). Hence, I categorized words in text reviews into "food," "service," "social," "expenditure," and "miscellany" categories and tested whether these topics affect the consumer sentiment indices:

Hypothesis 2 *H_2: The "food," "service," "social," and "expenditure" categories can affect consumers' positive and negative sentiment indices.*

In hypothesis H_2, I am concerned with whether these topic groups had significant impacts on the positive and negative sentiment indices. Moreover, I were interested in which topics had larger impacts on sentiments. I have demonstrated the advantages in using a fixed-effects model for the first estimating equation and used this method again:

$$SPSI_{it} = \gamma_0 + \gamma_1 Food_{it} + \gamma_2 Service_{it} + \gamma_3 Expenditure_{it} + \gamma_4 Social_{it} + \sigma_i + \epsilon_{it}, \tag{1.2}$$

[6] The main reason to standardize is that I were interested in comparing the impact of the positive and negative sentiment indices on ratings. Standardization made such a comparison plausible. See Cheadle et al. (2003); Marty et al. (2017); Garneau et al. (2018) for the standardization procedure and examples using standardized variables.

$$SNSI_{it} = \gamma_0' + \gamma_1' Food_{it} + \gamma_2' Service_{it} + \gamma_3' Expenditure_{it} + \gamma_4' Social_{it} + \sigma_i + \epsilon_{it}, \tag{1.3}$$

where $SPSI_{it}$ and $SNSI_{it}$ were the standardized positive and negative sentiment indices of restaurant i at month t, respectively; $Food$, $Service$, $Expenditure$, and $Social$ were fractions of the word count in that category compared to the word count in all categories;[7] σ_i was the restaurant fixed effect for each restaurant i; and ϵ_{it} was the disturbance term.

1.3 Results

Summary statistics

After data processing and textual analysis, my dataset includes the following variables: Yelp business ID, business name, date of review, star-rating of the restaurant, positive sentiment index, negative sentiment index, word count per review, and fractions for each category of words in review.

Table 1.2: Summary statistics. Yelp rating is a numerical scale from one-star to five-star with whole star increment. Values in sentiment analysis and semi-automatic content analysis are percentage.

Variables	Mean	Std. Dev.	Max	Min	Observations
		Yelp Review			
Rating	3.7	0.8	5.0	1.0	20,864
Word Count	129.9	74.4	955.0	3.0	20,864
		Sentiment Analysis			
Pos. Sentiment Index	6.7	2.5	37.5	0	20,864
Neg. Sentiment Index	1.1	0.9	25.0	0	20,864
		Semi-automatic Content Analysis			
Food	0.37	0.11	1.00	0	20,860
Service	0.13	0.08	1.00	0	20,860
Expenditure	0.02	0.03	0.51	0	20,860
Social	0.03	0.03	0.50	0	20,860
Miscellany	0.45	0.11	1.00	0	20,860

Notes. Positive (Negative) Sentiment Index of a review was defined as the fraction of the numbers of positive (negative) sentiment words in the review to the numbers of the word count of the review. The value of topics (Food, Service, Expenditure, Social, and Miscellany) in semi-automatic content analysis were defined as the fraction of high-frequency topic words in that category to the total high-frequency word count across all the topics.

Table 1.2 reports summary statistics for the panel dataset. From the table, I observed that (1) the mean rating was 3.7, with a standard deviation of 0.8, indicating that the average rating was over three and half stars and the one standard deviation change of rating was from three stars to four and half stars. The rating distribution was skewed toward five stars, which meant most consumers were

[7] I excluded "miscellany" category to avoid multicollinearity because the total share of the five categories is equal to 1.

satisfied with their dining experience. (2) The average word count was 130.0, with a standard deviation of 72.3. One standard deviation change from the mean was from 57.7 words to 202.3 words. Most reviews were between one or two short paragraphs. (3) From the sentiment analysis, the mean positive sentiment index was 6.7, with a standard deviation of 2.5, and the mean negative sentiment index was 1.1, with a standard deviation of 0.9. This is consistent with previous literature that positive words were much more likely to be used than negative words when describing food-related sentiments (Gmuer et al., 2015). The positive sentiment index exhibited more variation than did the negative sentiment index. (4) From semi-automatic content analysis, the mean of "food" was 0.37, which meant that, on average, food-related words made up 37% of the most frequently used words in all five categories. After "food", the order of mean from high to low was "service" (0.13), "social" (0.03), and "expenditure" (0.02) categories (Since "miscellany" was not a variable of interest, I did not include that category when ranking the means).

Relationships between consumer rating and sentiment indices

Table 1.3: Estimation results for equation (1). Yelp rating is a numerical scale from one-star to five-star with whole star increment. Std. Pos. (Neg.) Sentiment Index is percentage.

Variables	(1) Rating	(2) Rating	(3) Rating	(4) Rating
Std. Pos. Sentiment Index	0.649***	0.478***	0.436***	0.435***
	(0.00973)	(0.00924)	(0.00871)	(0.00880)
Std. Neg. Sentiment Index		-0.564***	-0.452***	-0.454***
		(0.00861)	(0.00812)	(0.00811)
Constant	3.785***	3.792***	3.786***	3.836***
	(0.00531)	(0.00484)	(0.00438)	(0.0116)
Observations	20,864	20,864	20,864	20,864
R-squared	0.176	0.316	0.450	0.454
Year FE	NO	NO	NO	YES
Restaurant FE	NO	NO	YES	YES

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Notes. Std. Pos. (Neg.) Sentiment Index is the fraction of the numbers of positive (negative) sentiment words to the numbers of total word count and then normalized by z-score. Year FE is the year fixed-effect which was used to control for some yearly time-varying factors (yearly-inflation). Restaurant FE is the restaurant fixed-effect which was used to control for time-invariant restaurant factors (restaurant locations, cuisine style, and etc.)

Table 1.3 summarizes the estimation results of equation 1.1, which were used to test $Hypothesis 1$, under different specifications. In column (1), I included only the standardized positive sentiment index (SPSI) as an explanatory variable for my baseline specification. The results suggested that SPSI was positively related to the restaurant rating, consistent with hypothesis H_1.

I included the standardized negative sentiment index (SNSI) in the second specification and reported the results in column (2) of table 1.3. There were three findings worth noting under this specification: First, the results suggested that SPSI (SNSI) was positively (negatively) related to the restaurant rating. Second, the point estimates of SPSI (β_1) were reduced significantly compared to the results in column (1), implying that SNSI should be included in the estimating equation and that failing to do so will incur bias in the estimated β_1. Third, by adding the SNSI variable, the R^2 was improved from 0.176 to 0.316, which suggested that SNSI could also explain a significant portion of the variation in the outcome variable. Overall, about 32% of variation in the restaurant ratings could be explained by the combination of the two sentiment indices.

By adding restaurant fixed effects, α_i, to the second specification, I had created a standard fixed-effects model. Restaurant fixed effects, α_i, controlled for any restaurant characteristics that did not change over time (e.g., restaurant location). I expected that the point estimates for SPSI and SNSI (β_1 and β_2) could change after adding fixed effects as they reduced the omitted variable bias in those estimates. Column (3) of table 1.3 shows the estimation results. I observed a mild change in both β_1 and β_2: Compared to the estimates in column (2), both β_1 and β_2 were reduced by roughly 10%, suggesting that the point estimates in column (2) could be biased. Moreover, the R^2 increased by 0.134 ($0.45 - 0.316$), which was the additional variation of the dependent variable explained by the restaurant fixed effects. This also implied that the sentiment indices explained a much higher share of the variation (31.6%) compared with time-invariant factors (13.4%).

Finally, as a robustness check, I added the year fixed effect α_T to the estimating equation. This fixed effect controlled for some time-varying factors (yearly price inflation for example) in the disturbance term that could lead to changes in β_1 and β_2. From column (4) of table 1.3, I found that after adding year fixed effects, estimated coefficients of SPSI and SNSI changed by less than 1%. This indicates that my results are robust. Second, the R^2 was changed by 1% after I included year fixed effect in my model. Overall, results from column (4) suggested that results from the standard fixed-effects model (column 3) were robust, making it my preferred specification.

Overall, the results from the standard fixed-effects model had four main implications: First, SPSI (SNSI) was significantly ($p < 0.01$) positively (negatively) related to rating, consistent with hypothesis H_1.

Second, the impacts of positive and negative words on ratings were roughly the same. My results suggested that a 1-standard-deviation increase in the SPSI and SNSI will lead to a 0.436 unit (that is, roughly half a star) increase and a 0.454 unit (again, roughly half a star) decrease in rating, respectively. From the summary statistics, I found that consumers tended to use more positive than negative words in reviews, but more positive words did not translate to a higher impact of positive sentiments on ratings.

Third, the R^2 reports that 45% variations in ratings were explained by SPSI, SNSI, and the restaurant fixed effect, of which 32% was explained by sentiments and 13% was explained by restaurant fixed effect. Hence, in my estimation, sentiment index explained most of the rating variations that was explained by the model. It meant that consumer rating was highly related to consumer sentiments instead of other time-invariant restaurant factors.

Which topics affect consumer sentiment word usage most?

Table 1.4: Estimation results for equation 1.2 and 1.3. Std. Pos. (Neg.) Sentiment Index is percentage. Topics are percentages.

	(1)	(2)
Variables	Std. Pos. Sentiment Index	Std. Neg. Sentiment Index
Food	0.747***	-0.0829**
	(0.0375)	(0.0410)
Service	1.493***	0.528***
	(0.0534)	(0.0584)
Social	-0.0176	-0.224*
	(0.124)	(0.136)
Expenditure	0.612***	0.271**
	(0.116)	(0.127)
Constant	-0.547***	-0.00584
	(0.0194)	(0.0212)
Observations	20,860	20,860
R-squared	0.136	0.103
Restaurant FE	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes. Std. Pos. (Neg.) Sentiment Index is the fraction of the numbers of positive (negative) sentiment words to the numbers of total word count and then normalized by z-score. The value of topics (Food, Service, Expenditure, Social, and Miscellany) were defined as the fraction of high-frequency topic words in that category to the total high-frequency word count across all the topics. Restaurant FE is the restaurant fixed-effect which was used to control for time-invariant restaurant factors (restaurant locations, cuisine style, and etc.)

Table 1.4 shows that the R squares of the equation (2) and (3) were both close to 0.1. This meant that around 10% variance were explained. There were two implications: 1) since only 10 percent of variations were explained by the topic variables, predictions using these variables alone would not be accurate. 2) The statistical significance of those topic variables indicated that a predictive model could include such variables to improve accuracy of predictions. Second, I can observe that "food," "service," and "expenditure" affected both consumer positive and negative sentiment word usage significantly but "social" did not affect SPSI significantly. I rejected hypothesis H_2, which assumed that all topics can affect consumer sentiment word usage. Besides qualitatively analyzing which factors had impacts on sentiment word usage, I were also interested in which topics impacted the magnitude of rankings.

In table 1.4, the coefficient before food meant that if we replaced a review that only contained "miscellany" topic with a review that only contained "food" topic, the SPSI will increase 0.747 and SNSI will decrease -0.0829. Also, the coefficient before service meant that if we replaced a review that only contained "miscellany" topic with a review that only contained "service" topic, the SPSI will increase 1.493 and SNSI will increase 0.528. I am able to observe that all the coefficients were the topic SPSI (SNSI) relative to "miscellany" topic. Since the relative SPSI (SNSI) were less intuitive, we "recovery" them to the absolute topic SPSI (SNSI) by adding the SPSI (SNSI) of miscellany topic.

Topic-specific SPSI and SNSI, which I defined as the SPSI or SNSI when reviews contained only one topic ("food," "service," "expenditure," "social," and "miscellany") are shown in table 4. When we calculated topic-specific SPSI (SNSI), we took the "miscellany" SPSI (SNSI) as my baseline, since the "miscellany" SPSI (SNSI) was the SPSI (SNSI) when reviews did not contain any of the other four topics. As long as we knew the "miscellany" SPSI (SNSI) and the SPSI (SNSI) changes when reviews changed from "miscellany" to one of the other four topics, we can obtain the other four topic-specific SPSI (SNSI).

Topic-specific SPSI and SNSI are found in table 1.4. By adding the constant term (γ_0)[8] to each topic point estimate in column (1),[9] we got the topic-specific SPSI. For example, the "food" SPSI was 0.2 $(0.747 + (-0.547))$, which suggested that, on average, if a review only contained food-related words, the SPSI of the review was 0.2. Topic-specific SNSI was derived in the same way. Summarizing the results from column (1) and (2), the SPSI ranked by topic from high to low was "service," "food," "expenditure," and "social." Topic SNSI ranked from high to low was "service," "expenditure," "food," and "social."

There are three main implications: First, service affected consumer's sentiment word usage the most when they were eating out. "Service" had the highest SPSI and SNSI among the four topics; reviews focused on service contained a higher proportion of sentiment-related words (both positive and negative) than other topics. One explanation for this is that restaurant service reflected interactions between consumers and restaurant staff (e.g., waitstaff, managers). I found that this kind of communication was more likely to stir up frequent sentiment word usage than other types of interactions (e.g., with food). However, the "social" category—which also described communications—did not have as much of an impact as "service." One possible explanation is that service-related reviews described

[8]The constant term γ_0 meant the SPSI when reviews contained only "miscellany"-related words.

[9]The estimated coefficient of each topic in column (1) meant the SPSI changes if reviews change from only containing "miscellany"-related words to reviews containing only words related to the specific topic.

interactions between consumers and restaurant staff, while social-related reviews described interactions among consumers. Since my data are restaurant reviews, consumers tended to only report restaurant-related communications rather than their own social activities. If future scholars are interested in what consumers said about their own social activities, Vidal et al. (2015, 2016) provides an example, using social media data (Twitter) to explore what people said in four dining circumstances (breakfast, lunch, dinner, snack).

Second, social-related words in reviews were used for social context description. "Social" had the lowest SPSI and SNSI. This suggested that reviews focused on social aspects contained fewer sentiment words than other topics. One plausible explanation was that when consumers mentioned social words in their reviews, they were describing social context rather than expressing themselves sentimentally. For example, one kind of common review involving social-related words was "...We always come here for family birthdays" which had no sentiment content.

Third, compared with the "expenditure" topic, consumers used more positive sentiment words and less negative sentiment words about "food." The SPSI of "food" was higher than that of "expenditure" and the SNSI of "food" was lower than that of "expenditure." Consumers were probably more satisfied with the "food" topic compared with the "expenditure" topic.

Which topics affect consumer rating most?

From the summary statistics in table 1.2, we knew that consumers were satisfied overall with their dining experience (i.e., the positive sentiment index was higher than the negative sentiment index and average rating was skewed to five stars). But we did not know which topics were most pleasing to consumers relative to others..

Next, I considered the topic-specific rating. I plugged the topic-specific SPSI and SNSI into equation (1.1) to obtain the topic-specific ratings: From high to low these were "service" (3.95), "food" (3.91), "expenditure" (3.69), "social" (3.56), and "miscellany" (3.55). These ratings indicated the average rating for reviews containing only words related to each category. There were two implications: First, "service" and "food" received ratings above the average sample rating (3.7). Topic-specific ratings of "expenditure" and "social" were lower than the average sample rating, suggesting that restaurants that want to improve their business ratings should focus on expenditure and social aspects. We also knew that "social"-related words did not contain as much sentiments relative to "expenditure"-related words.

Second, consumers felt most satisfied with "service," which received the highest rating among all topics. Consumers used the most frequent positive and negative sentiment words toward restaurant services, but positive words dominate negative words. It is worth clarifying that a consumer did not experience both positive and negative service experiences at the same dining experience. One possibility is that some consumers felt good about a restaurant, while others felt bad about it, but a more plausible scenario is that some restaurants received more positive sentiment words about their service while others received more negative sentiment words about theirs.

1.4 Discussion

I found that the average rating was skewed to five stars, suggesting that consumers were overall satisfied with their dining experiences. This is consistent with previous literature. For example, most emotions expressed in food-related tweets, either through words or emoticons and emoji, were positive rather than negative (Vidal et al., 2016). Even if there were more positive sentiment words in consumer reviews, the magnitude of the impact of positive and negative sentiment indices on reviews' ratings was almost the same. But still we cannot rule out a possibility that consumers who have an extremely good dining experience are more ready to leave reviews than consumers who have an extremely bad dining experience. The answer to this question was not obvious and cannot be solved by my dataset.

Restaurant ratings were positively related to SPSI and negatively related to SNSI, consistent with Canetti et al. (2002); Macht (2008); Edwards et al. (2013). This finding suggested that the sentiment index system (SPSI and SNSI) used in this case study was a good index to describe overall consumer satisfaction. Evidence from subsection 3.2 analysis also supported my argument. From the regression analysis, a considerable portion of rating variations (32%) was explained by sentiment indices. Future researchers can use the SPSI and SNSI sentiment index as one of predictors to predict consumer ratings in many food-related circumstances, including dining and grocery shopping. Evidence from the field of computer science research (Lei et al., 2016) supported adding sentiment factors to rating prediction algorithms to improve the accuracy of rating predictions.

When talking about "service topic", consumers used more sentiment words than other topics. I speculated that service involves interpersonal interactions, which can evoke strong sentiment. This finding is in agreement with Walls et al. (2011), who studied the luxury hotel industry and find that human interaction can affect customers' experiences. Even if "service" were linked with the most frequent sentiment word usage (both positive and negative), "service" still received the highest rating among all topics. All of the above findings led to one important point: service is important. However, restaurant owners must also consider that excessive service may also change consumer attitudes for the worse (Lin et al., 2019).

Social-related words had the least associated sentiment word usage (either positive or negative), compared with other topics. The purpose of online review platforms is to describe and evaluate a product or service. Reviewers' focuses were on product or service rather than other "external" things. In contrast, when posting on social media platforms, consumers' usually express their thoughts and feelings and talk about their private lives. For instance, Vidal et al. found that in food related posts from Twitter people frequently mentioned social contexts, especially referencing to their friends and family. Furthermore, when these social contexts were mentioned, the topic of the tweets focused on the user's social life and positive sentiments were usually expressed. As the audience for social media posts are usually friends and family, users are expected to be more likely to share their sentiments in these platforms. Since the audience of online reviews were restaurant owners and other consumers, consumers were less likely to expose their private lives. If future research can connect consumers' social media accounts with their online review platform accounts, this unique dataset could provide more comprehensive information about consumer behaviors on different platforms, thereby improving predictions of consumer behavior.

1.5 Limitations

In my case study, I manually categorized the most frequently used 300 words into five topics and calculated the fraction of each topic for each review. The top 300 words may not be enough to capture all consumer-concerned topics. The manual categorizing method limited my capacity to include more words in my topics. However, there is a trend toward using machine learning methods to categorize topics. Heng et al. (2018) used Latent Dirichlet Allocation (LDA) to automatically categorize coffee-shopping reviews from Amazon.com and found that service, physical, feature, flavor, and subjective expression can affect consumers' purchasing decisions. Future scholars can exploit the big-data advantage of online reviews to train their text-categorizing algorithms and categorize text reviews automatically,which will allow the word coverage of each topic to increase exponentially.

Data availability is an issue worth mentioning. For the research, I used Yelp data from the Yelp Dataset Challenge program, which gave researchers open access to Yelp review data. However, review platforms do not commonly provide their review data to researchers. In most situations, researchers need to collect these data by themselves. One common technique to collect data from websites is to use the web-scraping technique. Data displayed on websites can only be viewed by visiting its URL (uniform resource locator). Web-scraping is a kind of precoded software that can visit a particular website's URL and copy data from it to the user's local storage. Researchers can compile a customized web-scraping software by using Python or use a precoded package from R (Munzert et al., 2014).

The text analysis method I used in this paper was the semi-automatic content analysis method. It was different from a simple word count analysis, which calculated individual word counts in text data and had been deemed "uninformative at best and misleading at worst" (Vidal et al., 2015, p. 67). Vidal et al. (2015) recommended content analysis to analyze text data from Twitter, but this method requires a considerable time for researchers to analyze each text and extract its meaning in context. Hence, I used a compromise method that aggregated word count by topics. I was only concerned with how topic-related words affected consumer sentiment word usage, rather than with the impacts of individual words. For this purpose, my method was good enough.

It is important to note that after tokenization and stop word removal, words were considered out of their original contexts. There was a risk that the original meaning of the review may be misinterpreted. For example, if a review says “did not love the food”, after tokenization and stop word removal, only the words “love” and “food” would be associated to that review. The next question would be how serious this problem was in my analysis? The developers of the LIWC software discussed this situation based on probabilistic models of language use (https://liwc.wpengine.com/how-it-works/). Their conclusion was, if the lengths of the texts are over 50 words, even if the misclassification problem does exist, it will not impair conclusions to a serious degree. Since the average word count of Yelp text review was about 130 words, I believed that the impact of misclassification problem was very mild.

In sentiment analysis, the dictionary included not only emotion words but also consumer's affective responses toward their dining experience. Actually, the underlying assumption being made was that these affective words can reflect consumer's emotions indirectly. For example, the word “delicious” was assigned as positive sentiment in the LIWC. The underlying assumption was that when consumers wrote such a word in their review, they were likely to feel positive of their food.

1.6 Conclusion

I found that (1) consumers used more positive sentiment words than negative sentiment words in their Yelp restaurant text reviews. (2) The proportion of positive sentiment words in reviews was positively related to ratings; the proportion of negative sentiment words in reviews was negatively related to ratings. (3) Relative to food, consumers used more sentiment words (both positive and negative) when they were discussing restaurant service. (4) Consumers used the least sentiment words when they were discussing social-related topics compared with the other topics (food, service and expenditure). (5) Consumers rated restaurant service higher than they rated food.

My immediate contribution to sensory and consumer science study is providing a new insight to food-related consumer sentiment research by using online review data. Also, I found that online review data can provide comprehensive information for consumer behavior research. Admittedly, the 300 most-frequent words may not fully cover all consumer concerns, and the semi-automatic content analysis method may lose some estimation precision. However, as I claimed, my primary purpose in this paper was to provide insights into consumer sentiment studies from a new angle. For future researchers, it would be beneficial to explore more advance text analysis method to analyze online review data.

Chapter 2: Two-sided online review manipulation: sellers and platforms

2.1 Introduction

In the past three decades, e-commerce companies have evolved from small garage companies to business giants in this world. As of the first quarter of 2020, Amazon joins the trillion-dollar club, Yelp has more than 35 million average monthly mobile application unique users and 211 million cumulative reviews worldwide, and Expedia reaches 2.2 billion revenue. These fast-growing companies not only provide convenient purchasing channels for consumers, more importantly, but they also serve as consumers' opinion communication platforms by allowing them to leave reviews. Online reviews have become one of the most important information sources for consumers' purchasing decisions (Anderson and Magruder, 2012). Many consumers rely heavily on online reviews to guide their online purchasing decisions. Especially as spending habits change, as during the COVID-19 period, online reviews gain importance.

When product quality is unknown, online reviews can provide helpful guidance for consumers. Economists have long studied the relationship between experience goods and consumer information (Akerlof, 1970; Nelson, 1970). Some of the first online reviews appeared on eBay and research has shown that there are small price effects due to the reviews of the seller (Melnik and Alm, 2002), and these reviews are more important for inexperienced sellers (Livingston, 2005). Research also shows that online platforms have an incentive to promote high-quality sellers (Nosko and Tadelis, 2015; Jaffe et al., 2017).

Since online reviews gain huge impacts on consumer purchasing behaviors, review incentives begin to emerge. For example, an extra half-star increment at Yelp can increase restaurants up to 19% more reservations (Anderson and Magruder, 2012); also for Yelp, an extra one-star increment can lead to 5%-9% revenue increment. However, due to this attractive lure, businesses begin to manipulate with reviews to maximize their benefit. These manipulation behaviors can cause consumers' welfare loss, consumers' trust crisis (Anderson and Magruder, 2012), and in the long run, online review industry shrink (Luca and Zervas, 2016). Hence, the correctness and accuracy of online information used by consumers are critical.

Results of the existing literature about review manipulation are mixed. Some research shows that there are widespread fake reviews in online platforms (Mayzlin et al., 2014; Luca and Zervas, 2016). However, other research shows that there is no significant review manipulation in online platforms (Anderson and Magruder, 2012). To the best of my knowledge, the research that most closely matches my research was done by Mayzlin et al. (2014). In Mayzlin et al. (2014), by using a differences in differences approach, the authors examined the review difference of two hotel review platforms (TripAdvisor and Expedia) for different types of hotels. By assuming consumer homogeneity between the two platforms and no fakes reviews at Expedia, the authors assume that the rating difference between the two platforms comes from fake reviews posted by sellers at TripAdvisor. However, in my paper, I provide an alternative explanation for the rating difference between review platforms. I assume review incentives are not necessarily limited to fake reviews posted by sellers. Review platforms also have motivations and measures to change the incentives or even possibly manipulate reviews.

In this paper, I find that review platforms influence online reviews by affecting the writing behavior of consumers. To be specific, I compare the restaurant online review differences of Las Vegas, Nevada, between two online review platforms, Yelp and another industry-leading restaurant review platform. I find that the average rating of the industry-leading platform is a half star higher than Yelp for the same restaurants. By using the fixed-effects model, I explain the half star difference by inferring that the industry-leading platform intentionally attracts satisfied consumers to leave positive reviews. The reason that platforms can attract positive reviews is that there are costs for consumers to leave reviews (e.g. time cost for writing text reviews) and only consumers with a negative experience have enough motivation to leave reviews. Hence, if platforms can reward users to leave reviews, those positive consumers would have enough motivation to write reviews.

My immediate contribution to the review platform literature is that I find a new alternative explanation for average rating difference across platforms. This finding is meaningful for the future supervision of online review platforms. Furthermore, my paper provides empirical evidence for the information design theory literature (Sobel, 2020).

2.2 Data

My dataset is constituted by two components: Yelp review data and an industry-leader online restaurant reservation platform(hereafter platform O) review data. This unique long panel dataset contains reviews from 2007 to 2017 of 257 common restaurants at Las Vegas, Nevada, in the two platforms. It contains 155,568 review observations with Yelp and 219,386 observations with platform O. In this dataset, I have seven variables including, user name, restaurant name, restaurant address, review date, numerical rating, text review, and word count of each text review. Please see appendix A. for data processing.

From table 2.1, we can first observe that platform O has more observations than Yelp (219,386 vs 155,568). This finding might be counter-intuitive, because anyone can leave reviews on Yelp but only

Table 2.1: Summary statistics

Variables	Obs.	Mean	Std. Dev.	Min	Max
	Yelp				
Rating	155,568	3.9	1.3	1	5
Word Count	155,568	123.5	117.2	1	982
	Platform O				
Rating	219,386	4.3	1.1	1	5
Word Count	219,386	48.4	44.4	0	401

consumers who booked and honoured their reservations can leave reviews on platform O. However, if we consider that platform O is able to encourage consumers to leave reviews by giving them rewards, this finding makes sense. Second, I find that platform O has a higher average rating than Yelp (4.3 vs 3.9). The standard deviation of ratings is slightly higher for Yelp reviews (1.3) than for reviews at platform O (1.1). For the same restaurants during the same time period, users rated restaurants higher in platform O than Yelp. Third, I observe that word count of text reviews at platform O are shorter than Yelp (48.4 vs 123.5). The one standard deviation of Yelp and platform O word count is from 6.3 to 240.7 and from 4 to 92.8 respectively. Also, the maximum word count at Yelp is more than twice at platform O (982 vs 401). In short, under the reward policy of platform O, platform O attracts more reviews than Yelp. These reviews have higher ratings and shorter content than Yelp.

2.3 Theory

In this section, I discuss the existence of platform differences and explain the mechanism of platform incentives. The average ratings (for the same restaurants) between the two platforms have a half star difference. There are three possible reasons for this rating difference. First, sellers posted fake reviews. Second, heterogeneous consumers between the two platforms. For example, consumers at platform O are on average nicer than Yelp. Third, review platforms change the consumers' threshold of leaving reviews.

In my model, I assume that (1) there is no restaurant manufactured fake reviews at platform O. I am able to make this assumption is because platform O only allows consumers who booked on platform O and honored the reservations to leave reviews. In other words, the cost for restaurants to post fake reviews are too high. (2) Also, there is no restaurant manufactured fake reviews at Yelp. The validity of this assumption comes from the fact that Yelp has an unique and powerful fake review filter algorithm. The algorithm uses state-of-the-art machine learning techniques to analyse numerous confidential user background data (e.g., user's review history, location, login history, etc.) to detect fake reviews[1]. It is important to notice that I cannot promise that there is not even a single fake review at Yelp. However,

[1]See Yao et al. (2018) for more details about Yelp review filter algorithm.

most manufactured fake reviews are eliminated by Yelp's filter.

Differences in reviews can be explained by consumers' incentives. (see appendix B for theoretical model details). For consumers, I assume that there is a cost for them to leave reviews. Consumers will only leave reviews when their benefit of leaving a review is higher than their cost. Also, I assume that consumers who have a good experience are less motivated to leave reviews than consumers who have a bad experience. If platforms notice this feature and want to increase the number of reviews, their best strategy is to attract those consumers who are satisfied with their dining to leave reviews. In order to stimulate these consumers to leave reviews, platforms need to send them rewards to leave reviews.

Review platforms have motivations to encourage consumers to leave reviews, since having more reviews is a good signal for online review platforms (e.g., increase platforms' advertisement premium and stock price). Furthermore, these additional reviews are more positive than negative. Platforms are able to attract more restaurants to contract with them by promoting restaurant average rating at their platforms. Even if there is a cost for such incentives (the need to pay rewards to consumers), they may be able to transfer these costs to restaurants partially or entirely. Hence, they may have enough motivation to change the consumers' incentives. However, it is worth noticing that platform O is able to give rewards to consumers while Yelp cannot. This is because platform O can identify consumers' identities through booking information and restaurant feedback. Without consumer identity verification, it is imaginable that the reward policy would attract speculators that did not dine at the restaurant.

2.4 Empirical Method

The empirical model estimates whether (1) consumers at the two platforms are homogeneous and (2) platform O lowers consumers' leaving review threshold. These results may give evidence towards the existence of different platform incentives at platform O.

I estimate the relationship of the numerical rating and word count of the text review between Yelp and platform O. Numerical rating and word count are the two most important components of a review. Numerical rating represents an overall dining evaluation, and word count is able to reflect more details about consumers' dining experience. Evidence of similar rating and word count between the two platforms is a signal of homogeneous consumers between the two platforms. In order to control other factors that may affect rating, I use standard fixed-effect model and month dummy variable to control impacts from restaurant(location, cuisine style, decoration, price, and etc.) and time. I use two dummy variables to test the intercept and slope differences between Yelp and platform O. The estimation equation is given by:

$$Rating_{it} = \beta_0 + \beta_1 Yelp + \beta_2 Wordcount_{it} + \beta_3 Yelp * Wordcount_{it} + \sum_t \gamma_t \sigma_t + \alpha_i + \epsilon_{it} \tag{2.1}$$

where $Rating_{it}$ is the average rating of restaurant i at month t, $Yelp$ is an indicator function that is equal to one if the observation is from Yelp; if $Yelp$=0, the observation is from platform O. $Wordcount_{it}$ is standardized average word count of restaurant i at month t.[2] σ_t is a month dummy variable, α_i is restaurant fixed-effect, and ϵ_{it} is the error term.

2.5 Results

Table 2.2 shows the estimation results from equation 2.1. First, we observe that the coefficient for $Yelp$ is -0.51 and significant at 1%. This implies that given a word count of zero, platform O has higher ratings than Yelp. The coefficient for $Wordcount$ is -0.39 and significant at 1%. Hence, the word count and rating are negatively correlated. The more contents consumers would like to write, the lower rating they would like to assign. This implies that bad mood normally motivates people to express more.

Table 2.2: Estimation result of equation 2.1

Variables	Rating
1_{yelp}	-0.510***
	(0.0190)
Wordcount	-0.390***
	(0.0183)
1_{yelp}*Wordcount	0.00488***
	(0.000302)
Constant	4.045***
	(0.0332)
Observations	23,028
No. of restaurants	254
R-squared	0.140
Restaurant FE	Yes
Month Dummy	Yes

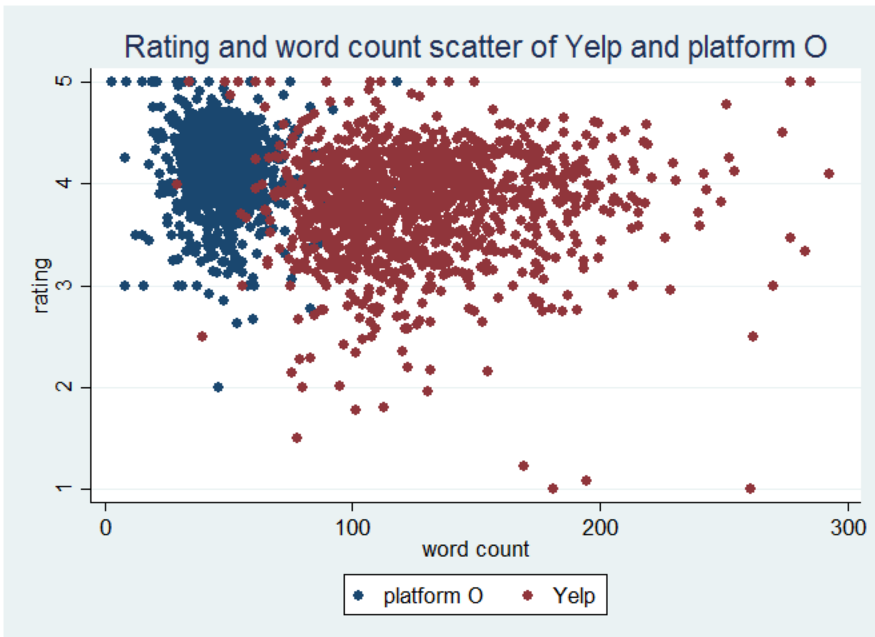
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficient for $Yelp * Wordcount$ is 0.00488 and significant at 1%. This intersection term tests the slope of rating and word count differences between Yelp and platform O. Its coefficient is positive and significant means that the slope of Yelp, with respect to the word count, is less steep than platform

[2] $Wordcount$ has been standardized using a Z score, such that $Wordcount = \frac{Word\ Count - 87.15}{62.79}$ since 87.15 is the mean and 62.79 is the standard deviation for the sample that contains both Yelp and OpenTable. Hence the -.39 implies that if the review is one word longer, the expected rating would change $-.39 * 1/62.79 = -0.006$.

O. However, it is small relative to the absolute value of coefficient for $Wordcount$. Hence, we can deem that the slopes of Yelp and platform O are similar enough that given any world count within the range of the sample, the expected review is higher on platform O than for Yelp. From table 2.1, we know that the average rating for Yelp and platform O is 3.9 and 4.3 respectively; the average word count is 123.5 and 48.4. From figure 2.1, we can see clearly that the two platforms attracted a different subset of consumers to leave reviews. Platform O attracts more satisfied consumers than Yelp.

Figure 2.1: Rating and word count scatter



I make two adjustments for better data display: (1)data is collapsed to yearly level. (2)6 outliers, whose word count over 300, are dropped.

The slope of Yelp is also negative(-0.39+0.00488). This result shows both consumers at the two platforms tend to write more words when they have bad dining experience (low rating). Hence, based on these evidence, I assume that consumers at the two platforms have other similar features (rating rubric, taste, and etc.). Furthermore, I conjecture that consumers are homogeneous between the two platforms. In short, we have evidence for homogeneous consumers and for the existence of varying platform incentives.

2.6 Conclusion

In this paper, I find that review platforms are able to affect consumer reviews by changing consumer review behaviors. This finding provides a new perspective for online review platforms. It is beneficial for future researchers to explore more possible ways of information manipulation, besides analyze fake reviews. Broadly, variety online review platforms and their different information display designs provide the potential to study the impacts of information display frames on consumer behaviours.

References

Akerlof, G. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Anderson, M. and Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989.

Campbell, A., Mayzlin, D., and Shin, J. (2017). Managing buzz. *The RAND Journal of Economics*, 48(1):203–229.

Canetti, L., Bachar, E., and Berry, E. M. (2002). Food and emotion. *Behavioural processes*, 60(2):157–164.

Carbon, K., Fujii, K., and Veerina, P. (2014). Applications of machine learning to predict yelp ratings.

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using z score transformation. *The Journal of molecular diagnostics*, 5(2):73–81.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Edwards, J. S., Hartwell, H. J., and Brown, L. (2013). The relationship between emotions, food consumption and meal acceptability when eating out of the home. *Food Quality and Preference*, 30(1):22–32.

Felbermayr, A. and Nanopoulos, A. (2016). The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing*, 36:60–76.

Garneau, N. L., Nuessle, T. M., Mendelsberg, B. J., Shepard, S., and Tucker, R. M. (2018). Sweet liker status in children and adults: Consequences for beverage intake in adults. *Food Quality and Preference*, 65:175–180.

Gmuer, A., Guth, J. N., Runte, M., and Siegrist, M. (2015). From emotion to language: Application of a systematic, linguistic-based approach to design a food-associated emotion lexicon. *Food Quality and Preference*, 40:77–86.

Heng, Y., Gao, Z., Jiang, Y., and Chen, X. (2018). Exploring hidden factors behind online food shopping from amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, 42:161–168.

Huang, J., Rogers, S., and Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo).*

Jaffe, S., Coles, P., Levitt, S., and Popov, I. (2017). Quality externalities on platforms: The case of airbnb.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage publications.

Ladhari, R., Brun, I., and Morales, M. (2008). Determinants of dining satisfaction and post-dining behavioral intentions. *International Journal of Hospitality Management*, 27(4):563–573.

Lei, X., Qian, X., and Zhao, G. (2016). Rating prediction based on social sentiment from textual reviews. *IEEE Transactions on Multimedia*, 18(9):1910–1921.

Lin, C.-F., Fu, C.-S., and Chen, Y.-T. (2019). Exploring customer perceptions toward different service volumes: An integration of means–end chain and balance theories. *Food quality and preference*, 73:86–96.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89.

Liu, Z. and Park, S. (2015). What makes a useful online review? implication for travel product websites. *Tourism Management*, 47:140–151.

Livingston, J. A. (2005). How valuable is a good reputation? a sample selection model of internet auctions. *The Review of Economics and Statistics*, 87(3):453–465.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):iv–vii, 3393–3672.

Macht, M. (2008). How emotions affect eating: a five-way model. *Appetite*, 50(1):1–11.

Marty, L., Chambaron, S., Bournez, M., Nicklaus, S., and Monnery-Patris, S. (2017). Comparison of implicit and explicit attitudes towards food between normal-and overweight french children. *Food quality and preference*, 60:145–153.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–55.

Melnik, M. I. and Alm, J. (2002). Does a seller's ecommerce reputation matter? evidence from ebay auctions. *The Journal of Industrial Economics*, 50(3):337–349.

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining.* John Wiley & Sons.

Nakayama, M. and Wan, Y. (2018). Is culture of origin associated with more expressions? an analysis of yelp reviews on japanese restaurants. *Tourism Management*, 66:329–338.

Nakayama, M. and Wan, Y. (2019). The cultural impact on social commerce: A sentiment analysis on yelp ethnic restaurant reviews. *Information and Management*, 56:271–279.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, 78(2):311–329.

Nosko, C. and Tadelis, S. (2015). The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennebaker, J. W. and Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, 10(6):601–626.

Piepho, H.-P. and Kalka, E. (2003). Threshold models with fixed and random effects for ordered categorical data. *Food Quality and Preference*, 14(5-6):343–357.

Schouteten, J. J., Verwaeren, J., Gellynck, X., and Almli, V. L. (2019). Comparing a standardized to a product-specific emoji list for evaluating food products by children. *Food quality and preference*, 72:86–97.

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, 128(3):907–947.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Vidal, L., Ares, G., and Jaeger, S. R. (2016). Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference*, 49:119–128.

Vidal, L., Ares, G., Machín, L., and Jaeger, S. R. (2015). Using twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations". *Food Quality and Preference*, 45:58–69.

Walls, A., Okumus, F., Wang, Y., and Kwun, D. J.-W. (2011). Understanding the consumer experience: An exploratory study of luxury hotels. *Journal of Hospitality Marketing & Management*, 20(2):166–197.

Wu, A. et al. (2017). Gender stereotype in academia: Evidence from economics job market rumors forum. Technical report.

Yao, Y., Angelov, I., Rasmus-Vorrath, J., Lee, M., and Engels, D. W. (2018). Yelp's review filtering algorithm. *SMU Data Science Review*, 1(3):3.

Yu, B., Zhou, J., Zhang, Y., and Cao, Y. (2017). Identifying restaurant features via sentiment analysis on yelp reviews. *arXiv preprint arXiv:1709.08698*.

Appendix A: data processing

It takes us three steps to construct the dataset. First, I construct the Yelp dataset. The Yelp review data comes from the Yelp dataset challenge program which provides free access to about 4.5 million local-business (restaurants, coffee shops, barbers, and etc.) and consumer reviews across 10 metropolitan areas in United States and Canada. This dataset contains comprehensive consumer (e.g., numerical rating, text reviews, and etc.) and business description variables (e.g., whether the business has parking, allows pets, has WiFi availability, and other 78 variables). I extract restaurant reviews of Las Vegas from these data to build the Yelp dataset. Due to two reasons, I choose Las Vegas as the sample: (1) unlike most other cities in the Yelp dataset challenge program, Las Vegas is a tourist city. Hence, most of the reviews on Las Vegas restaurants are left by tourists. Compared with tourists, local consumers have established their expectations about local restaurants, which make them less likely to leave reviews without undergoing extreme dining experiences. Therefore, reviews from tourists are more representative than local residents. (2) Processing the all 10 metropolitans outperforms my computer processing capacity.

Second, I use web-scraping technique to collect restaurant consumer review data of Las Vegas at platform O. I collect 6 variables including user name, restaurant name, restaurant address, numerical rating, text reviews and review date. Also, I dropped all reviews out of the range of January 2017 to December 2017.

Third, I merge the two sub-datasets by restaurant name and address. Before matching, I adjust all restaurant name spelling to lower case for higher successful matching probability. For unmatched restaurants, I manually check their restaurant name spelling and address. Finally, I obtained 257 matched restaurants from the two platforms.

Appendix B: A simple model

Model set up

I assume that there is one restaurant, n ($n \in \mathbb{Z}^+$, $n > 1$) identical consumers, and two online review platforms(platform A and platform B) in the economy. Further, I assume that platform A encourages consumers to leave reviews by sending them rewards. Platform B chooses not to send rewards to consumers.

I assume that consumers are able to obtain utility from their dining process. Before dining, all consumers share the same expected dining utility.

$$E(u(c)) = \bar{u} \tag{B.1}$$

here, c is consumer's dining quality; $u(\cdot)$ is consumer's utility function, $u(\cdot)$ satisfies $u'(\cdot) > 0$, $u''(\cdot) < 0$; $\bar{u}$ is consumer's expected dining utility.

After dining, consumers are able to observe the real dining quality. The quality is either good or bad. I assume that consumers with good dining quality would assign five stars to the restaurant and 1 star with bad dining quality. Also, I assume that consumers with bad dining quality are more likely to influence with reviews (leave long reviews) than consumers with good dining quality. It is because consumers are more emotionally sensitive to bad quality than good quality. Also, writing reviews is one of consumers' outlet for their accumulated emotions. Hence, the utilities of good and bad dining quality are:

$$u(c_{good}) = \bar{u} + \gamma \tag{B.2}$$

$$u(c_{bad}) = \bar{u} - \theta \tag{B.3}$$

where c_{good} is good dining quality; c_{bad} is bad dining quality; γ is the utility bonus of good quality; θ is the disutility of bad quality; since a bad quality dining has more impacts on consumers' dining utility than a good one, I assume $\theta > \gamma$.

Based on real dining quality and expected dining quality, consumers make their review writing decisions (leave reviews or not and length of text reviews). I assume that the more difference between consumers' expected dining quality and real dining quality, the more likely consumers would like to influence with reviews. The willingness to leave review function is:

$$W(c_{real}) = f(|u(c_{real}) - E(u(c))|) + m + \epsilon \tag{B.4}$$

here, $W(\cdot)$ is the final word count wrote by consumers; $f(\cdot)$ is the word count function, which is a monotonic increasing function of the difference between consumer real dining quality and expected dining quality($f'(\cdot) > 0$, assume there is $a, a > 0$, *subject to* $f(x) = 0$ if $x \leq a$; $f(x) > 0$ if $x > a$; also, $\gamma < a < \theta$); m is the extra words wrote by consumers who are motivated by the platform A's rewards(for consumers with platform B, $m = 0$); ϵ is the disturbance term, which follows standard normal distribution.

Consumers' review writing decisions

After substituting consumers' expected utility (equation B.1), good and bad dining utility (equation B.2 and B.3) into consumers' willingness to leave review function (equation B.4), we obtain consumers' review writing decisions. Below are the word counts that consumers left under good and bad dining quality respectively:

$$W(c_{bad}) = f(\theta) + m + \epsilon \tag{B.5}$$

$$W(c_{good}) = f(\gamma) + m + \epsilon \tag{B.6}$$

Figure B.1 shows consumers' different leaving review decisions between the two platforms. From figure B.1, I have two findings: first, for good dining quality, platform B users don't have enough motivations to leave reviews. While after receiving rewards, users from platform A would like to leave reviews. Since their experience are good, consumers wouldn't tend to influence with long reviews. Hence, users with platform A would assign five star ratings (assumed in model set up) and write short text reviews. Second, for bad dining quality, both of users on the two platforms would like to influence other consumers with long reviews. Reviews on the two platforms would be one star ratings with long text reviews. At the same time, text reviews at platform A are longer than platform B because users with platform A motivated by extra rewards.

We can observe that the difference of reviews between the two platforms depends on the ratio of good and bad consumer dining experiences. If there are more positive consumers than negative, platform A would have more reviews, higher ratings, and shorter text reviews than platform B. While if most of consumers feel negative about their dinners, platform A still has more reviews and higher ratings than platform B but with longer text reviews.

Figure B.1: Consumers' review writing decisions between the two platforms