HOW HOSPITALITY FIRM STRATEGIES AFFECT CONSUMER BIASES IN ONLINE REVIEWS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Fulfillment of the Requirements for the Degree of

Master of Science

by

Xinhua Wang

August 2021

ABSTRACT

Online reviews have become increasingly important to both consumers and businesses and, as a result, have attracted considerable research attention. However, all reviews are not created equal, as consumers may differ in their propensities to leave reviews, often as a function of their satisfaction. To ensure a more representative customer voice, companies often utilize different strategies to moderate the biases in online reviews. The strategies deployed by many hospitality firms differ dramatically in both how reviews are collected and where they are posted. This study investigates five review collection strategies of major hospitality companies and analyzes how each strategy affects review metrics (e.g., rating, length, and sentiment). We find that the effort required to post a review impacts review characteristics. We show that reviews collected through self-motivation methods tend to be lower-rated and longer, whereas reviews solicited from companies through post-stay emails tend to exhibit different characteristics. To measure the impact of the collection methods on review sentiment, we explore five different sentiment analysis methods and find results that are inconsistent both across the analysis methods and with other review metrics.

Biographical Sketch

Xinhua(Frances) Wang received her bachelor's degree from the University of Nevada, Las Vegas. During her years in Las Vegas, she explored different jobs in the restaurant, hotel, exhibition, casino, and travel industries, where she developed wonderful memories and made lifetime friends. In her senior year, she entered the inspiring world of research, where she began to wonder if the meaning of life lies not in having all the answers but in pursuing unanswerable questions with good company. From there, she decided to continue her journey by pursuing the research master's degree at Cornell University, and along the way, she has met many like-minded people who have encouraged her to write this thesis.

This paper is dedicated to my parents.

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my research supervisor, Dr. Chris Anderson, for giving me the opportunity to do research and providing invaluable guidance throughout this process. His dynamism, vision, sincerity, and motivation have been deeply inspiring. I would also like to thank him for this friendship, empathy, and great sense of humor. Without him, I would not have been able to finish this thesis, and I would not have had the courage to keep pursuing the research road. Besides my supervisor, I would also like to thank my thesis committee member: Dr. Helen Chun. She can always give me clear instructions of what I need to do next when I was lost.

I have been fortunate to have met many like-minded people at Cornell, and I really appreciate Cornell University for being such a generous and inspiring place that makes me feel the truth of "any person any study" every day. Thanks to Dr. Robert Kwortnik and Ellen Marsh for organizing so many wonderful Brown Bag Seminars where I had the chance to communicate with faculty and Ph.D. students, (and the yummy food was always a plus). Thanks to Dr. Sumanta Basu for lighting up my passion for statistics. Special thanks to Dr. Sachin Gupta for leading me into the world of quantitative marketing and giving me the chance to be his TA for hundreds of MBA students when I was not confident enough to take this role. I also sincerely appreciate Dr. Shawn Mankad for always being approachable and constantly providing help for my questions. All of these professors have set up the model for what kind of professor I want to be (if I have the chance) one day.

I would also like to sincerely thank my cohort fellows who have experienced ups and downs with me throughout this process and always make

me feel I am not alone. Thanks to Dr. (to be) Andrew Foley for always being a supportive and funny friend who has given my Master's life so much fun and inspiration. He always makes me feel that life is easy even when it is not. Thanks to Dr. (to be) Yue Liang for being my twisted sister (refer to Grey's Anatomy) and being my reality checker when I was too optimistic. Thanks to Jose Medina for being my proofreader and mental supporter. Thanks to Zihao Chen for being my patient debugger whenever I had a bug. Thanks to my UNLV squad who also came to Cornell and brought so much happiness to my life: Torry Sai, Pris Zhang, Sunny Wang (and her dog Sakura), and Selena Zeng. Thanks to my barbeque sister Ivana Liang for always being there for me no matter what.

Last but not least, my deepest gratitude goes to my parents. They are my twin pillars, without whom I could not stand. They are my ultimate inspiration, best friends, and role models. They never gave me any idea that I could not do whatever I wanted to do or be whomever I wanted to be. I can't express how lucky and proud I feel to be their daughter.

Contents

List of Tables

List of Figures

1 Introduction

Online reviews are becoming increasingly crucial for businesses. With the increasing number of such reviews, consumers have started to rely more on them to make purchase decisions. According to the 2019 Local Consumer Review Survey by BrightLocal, 82 percent of consumers read online reviews from local businesses, with 52 percent of people aged 18-54 reporting that they always read reviews. Most consumers read ten reviews before feeling confident that they can trust a company (BrightLocal, 2019). Apart from the crucial role that online reviews play in consumers' decision-making process, researchers have also proven that online reviews have a direct causal impact on business success. Even the tiniest rating change—for example, a half-star improvement on Yelp—can lead a restaurant to fill 30 to 49 percent more seats during peak hours (Anderson and Magruder, 2012).

However, online reviews might also raise some concerns. There is a systematic flaw that appears in almost all online reviews. Online review behavior is broadly understood within the literature to be a social exchange behavior (Wasko and Faraj, 2005; Liang et al., 2008; Lee et al., 2006; Kankanhalli et al., 2005). According to social exchange theory (SET), every social interaction is a consequence of benefits and costs (Cook et al., 2013). Consumers are more likely to be involved in online sharing behavior when its perceived benefits outweigh the perceived costs. Consumers may perceive the benefits and costs of posting behavior differently due to their different experiences and habits, such as familiarity with a platform (Min Kim et al., 2020; Schoenmueller et al., 2020; Han and Anderson, 2020). Thus, different groups of consumers have different propensities to leave reviews (Han and Anderson, 2020), which causes most on-

line reviews to suffer from selection bias. Schoenmueller et al. (2020) compared 25 major online review websites and found polarity and a positive imbalance on most platforms. Therefore, online reviews tend to overrepresent the most extreme views (Klein et al., 2018).

To provide a more accurate image of consumer voices, firms apply different strategies to encourage online review posting behavior across different kinds of customers, using different methods to increase motivations or decrease costs. Many online review platforms such as TripAdvisor utilize the incentive hierarchy system, where consumers can accumulate points by posting reviews and show their contribution status on their profiles. This system helps consumers internalize at least some of the benefits of sharing opinions (Liu et al., 2016). Most platforms utilize email invitations to encourage consumers to leave reviews, as these invitations can be considered a personal connection with the company and can reduce the cost of actually going to a platform to leave the reviews. Monetary and prosocial incentives are also widely used strategies among firms. The online review system on Glassdoor uses a "give-to-get" policy that provides strong incentives for consumers to leave reviews: after viewing a certain number of reviews, users are asked to submit their own experiences in the community to read more reviews (Klein et al., 2018).

Hospitality firms also utilize online reviews to help promote their businesses. Currently, there are three main kinds of sites that provide online hotel reviews: community-based sites such as TripAdvisor, transaction-based sites such as official brand websites and Expedia, and metasearch engines such as Google. Due to these sites' different business models, the online reviews on them serve different purposes. Community-based websites such as TripAdvi-

sor value both the quantitative and qualitative information in online reviews because they are the basis of a user-generated content business model (Miguéns et al., 2008). The more high-quality content supplied on a company's website, the more time that consumers spend surfing the website. This allows more advertisements to be displayed and offers more opportunities to reach consumers, so more revenues can be generated (Kumar and Benbasat, 2006). The goal of transaction-based sites is to help consumers make purchase decisions quickly and efficiently. Therefore, such websites want to have as many reviews as possible to reduce uncertainty in consumer decisions. The logic behind pursuing high volumes of online reviews is rooted in herding behavior and social impact theory. Knowing that peer-generated reviews are authentic makes consumers more likely to book because many other people have already experienced the product or service, so the risk of making the wrong decision is reduced (Banerjee, 1992). The mere existence of consumer opinions has an influence on other consumers, regardless of whether these opinions are positive or negative (Godes and Mayzlin, 2009; Xiong and Bharadwaj, 2014). Metasearch engines also rely on traffic, but instead of serving mainly as an information destination, they aggregate reviews from across various sources and direct customers to different websites. The presence of online reviews has been shown to positively influence consumers by improving customer perceptions of the usefulness and social presence of websites (Kumar and Benbasat, 2006). Reviews have the potential to attract consumer visits, increase website stickiness, and create a feeling of community among frequent shoppers (Mudambi and Schuff, 2010).

Hospitality firms utilize different strategies to collect and present online reviews to encourage review posting behavior. There are distinct differences across hotel brands in terms of where reviews collected from verified customers

are displayed. Among major hotel brands, some host their own review page on their official brand websites, while others cooperate with TripAdvisor directly and post all collected reviews to that site. Major hotel brands such as Marriott and Choice not only host review pages on their brand websites but also display reviewers' membership status next to their reviews. In contrast, some hotels such as Hilton and Wyndham display a few reviews from TripAdvisor on their own websites and direct all consumers to leave reviews on TripAdvisor. One of the reasons for this choice is the credibility of the third-party review website. After studying customer reviews of electronic products on both third-party review platforms and brand websites, Wu and Lin (2017) found that the same reviews posted on third-party review platforms are more credible to customers than those posted on brand websites. The authors also determined that the perceived helpfulness of reviews is higher on third-party websites than on brand websites. Beyond hotel brand websites, tourists also share their travel experiences through online travel agencies (OTAs) such as TripAdvisor (Guo et al., 2017; Liu et al., 2018; Kim and Hyun, 2021), Expedia (Xiang et al., 2015; Schoenmueller et al., 2020), and Yelp (Papathanassis and Knolle, 2011; Schoenmueller et al., 2020). In March 2021, TripAdvisor had over 570 million reviews of the world's leading hotels (DMR, 2021). Expedia and Booking.com are two OTA giants, dominating over 92 percent of the OTA market in the US (Research, 2019). In 2019, Expedia alone accounted for 35 percent of all OTA revenue worldwide (Statista, 2021).

The review collection process differs among hospitality firms. Since these firms utilize online reviews for different purposes, their goals in the collection of online reviews are different, which leads to different methods of soliciting online reviews. Figure 1 shows the simple template for submitting a review

through Expedia. Expedia sends consumers a link to review properties, and consumers go through a few rating options and then leave a review. Expedia does not require a minimum word count for reviews, meaning that consumers can submit an empty review. Figure 2 shows a template for submitting a review through a guest satisfaction survey (GSS). A GSS is created by hotel brands and usually contains a very long survey about the guest experience and then a section to leave a review. Brands such as Hilton and Wyndham do not have their own review section and require consumers to log onto TripAdvisor at the end and leave a review there. Figure 3 shows the template for submitting a review through TripAdvisor. TripAdvisor requires consumers to leave a rating and a review over 200 characters. Then, TripAdvisor has an optional rating section for brands to include additional questions.

For brand websites, taking Marriott as an example, consumers are sent an email soliciting their opinions regarding their stay after they check out. In a GSS, consumers are asked to fill out a survey and leave their reviews. Such surveys are only for company use and are not public, but ratings and reviews may be shared on the brand website. For OTA sites such as Expedia, reviews are also usually generated from invitation emails. Consumers who book through Expedia receive a short survey that invites them to rate and review their hotels. On TripAdvisor, there are three ways to post reviews. First, consumers can voluntarily go to TripAdvisor to leave their reviews; this kind of review, illustrated in Figure 4, is called a self-motivated (SM) review. Second, unlike Marriott, which has its own online review page on its brand platform, hotels such as Hilton encourage consumers to leave a review on TripAdvisor at the end of the GSS invitation. This kind of review appears on TripAdvisor labeled "in partnership with this brand", as shown in Figure 5. Third, although Marriott hotels collect

Figure 1: Expedia Review Collection Template

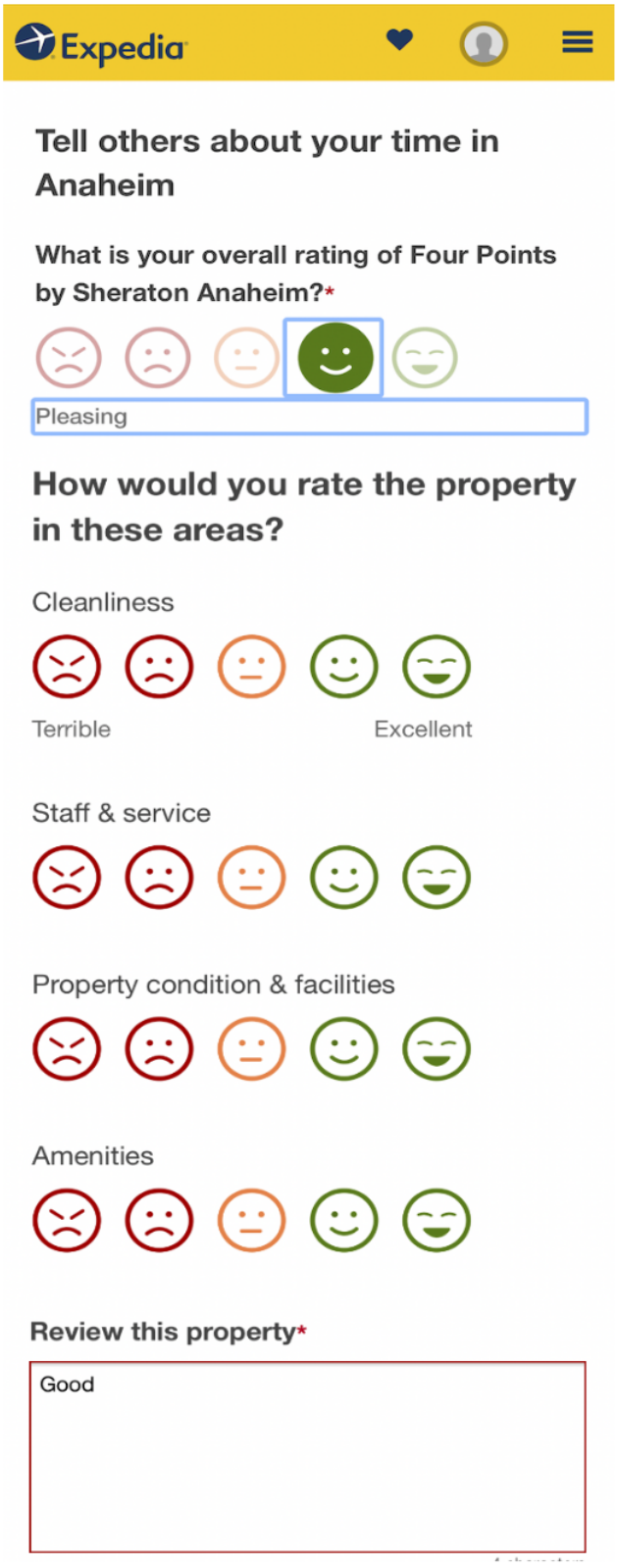


Figure 2: GSS Review Collection Template

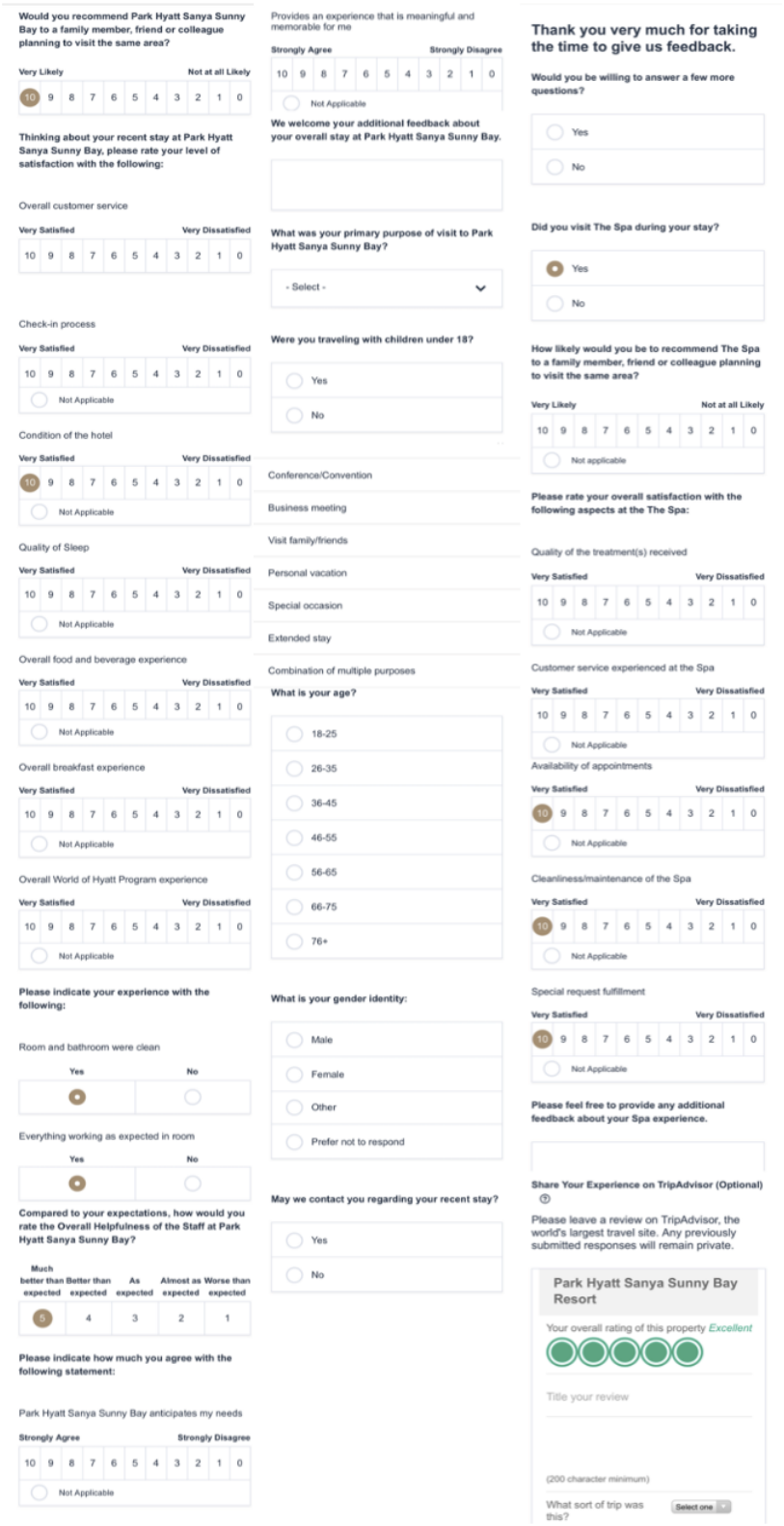
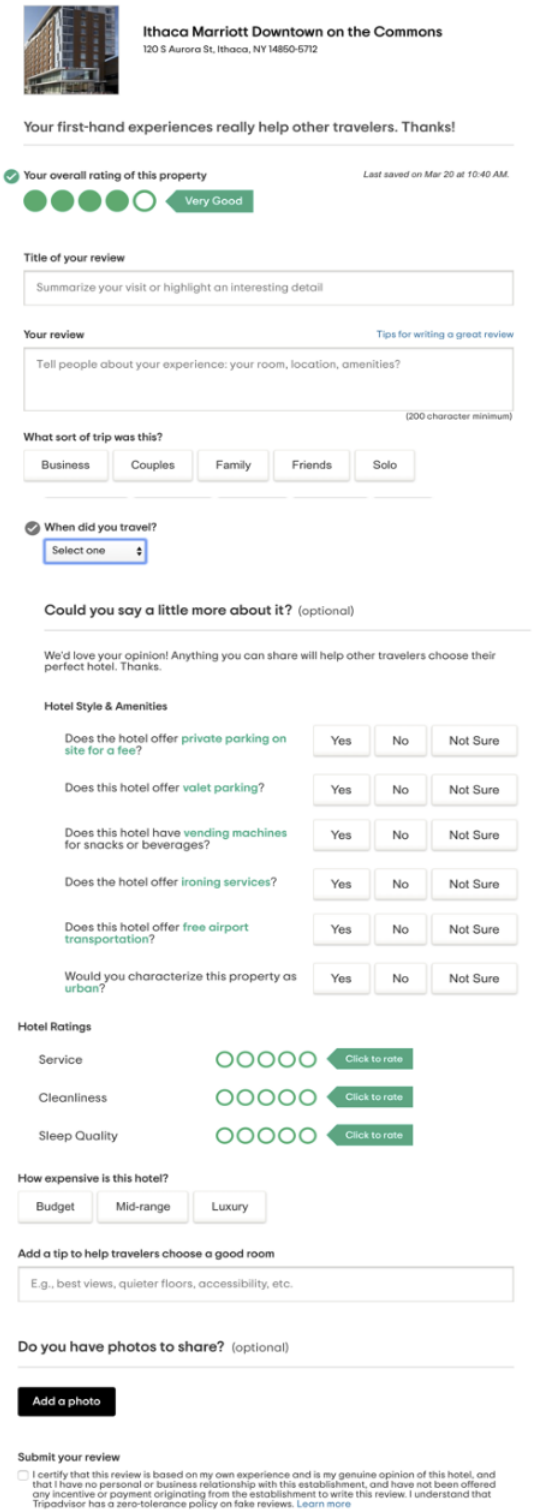


Figure 3: TripAdvisor Review Collection Template



reviews and post them on their own brand websites, some hotel owners still cooperate with TripAdvisor and use TripAdvisor's review collecting partner, Review Express (RE), to post reviews on TripAdvisor, again labeled "in partnership with this hotel", as shown in Figure 6. Only a small portion of big-brand hotel companies use this collection method. In this study, we dub these five collection methods Brand, Expedia, TA-GSS, TA-SM, and TA-RE, accordingly.

Figure 4: TA-SM Review

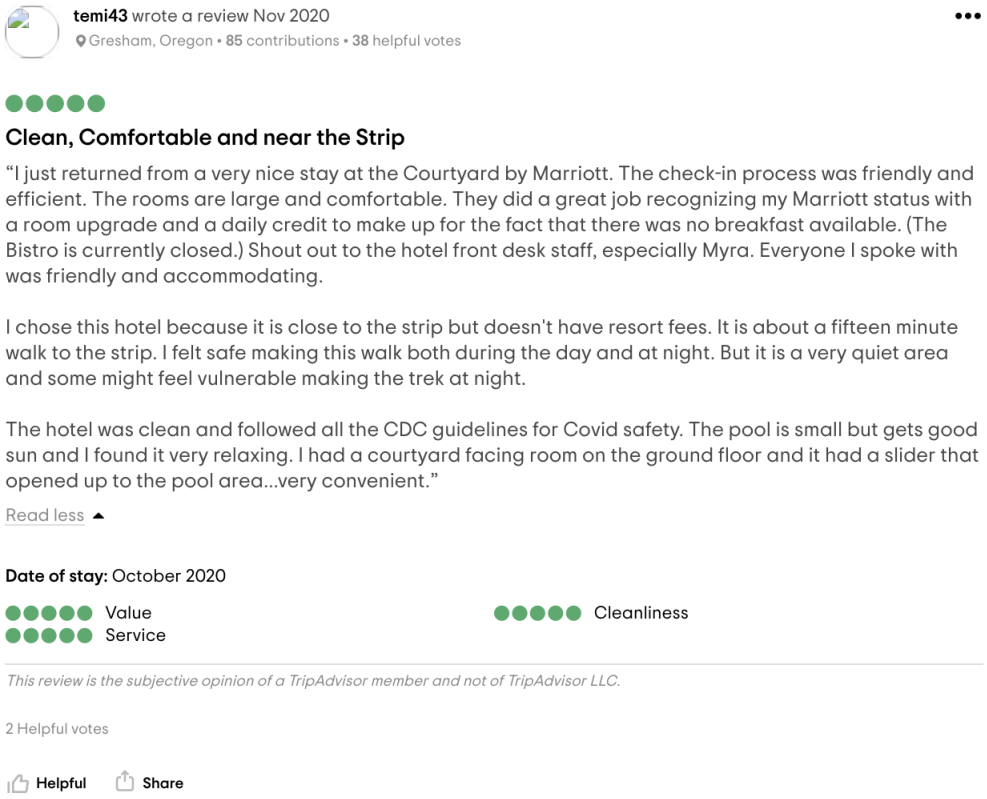


Figure 5: TA-GSS Review

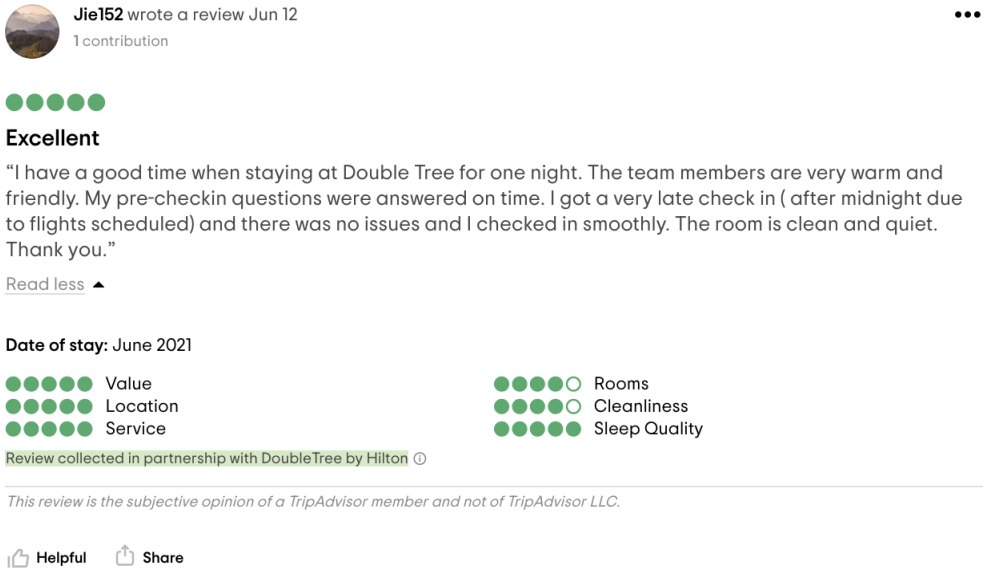


Figure 6: TA-RE Review

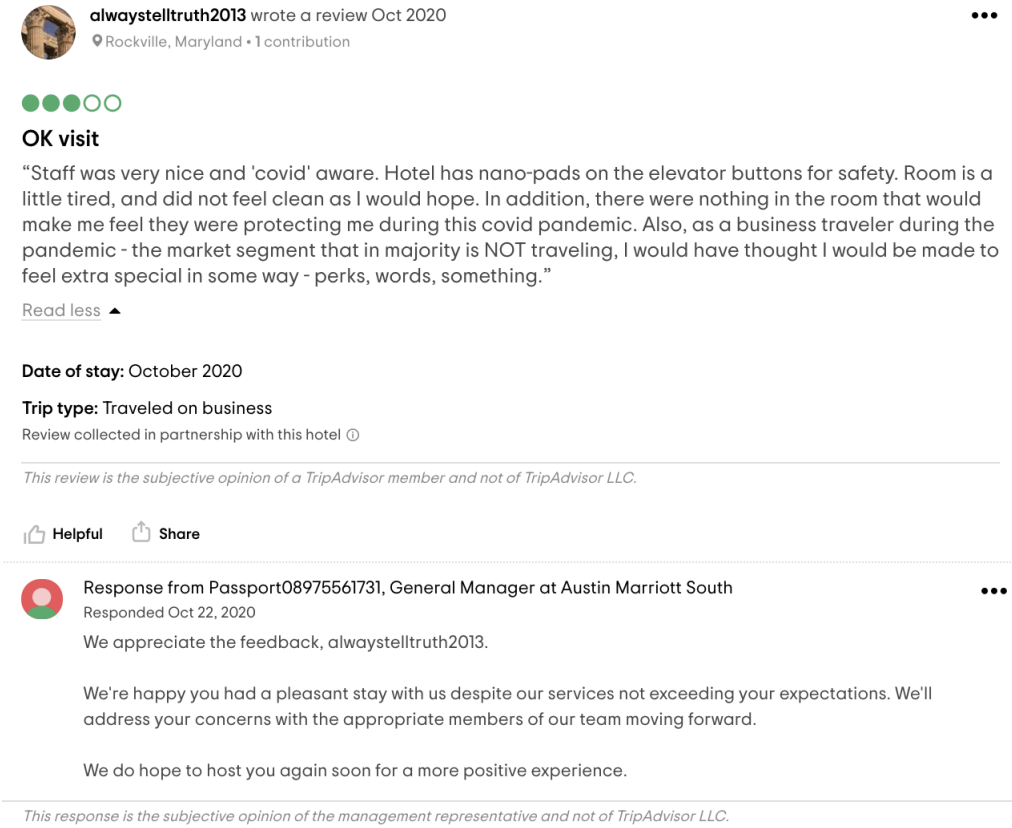


Table 1 presents the main differences in the review collection process among these five collection methods. In an effort to identify how different collection methods can affect review metrics, I collected review data from a series of hotel brands across different review platforms. The way that consumers are prompted to leave a review as well as the costs of submitting a review on a certain platform can influence different groups of customers and the contents of the reviews that customers write.

Table 1: Posting Cost Differences Between Collection Methods

	Cost	Brand	Expedia	TripAdvisor		
				GSS	SM	RE
Email Invitation	-	YES	YES	YES	NO	YES
Private Survey	+	YES	NO	YES	NO	OPT
Log-in	+	NO	NO	YES	YES	YES
Review Details	+	YES	NO	YES	YES	YES
Posting Costs Rank (Low to High)		2	1	4	5	3

In this study, I hope to answer the question of how firm collection strategies affect online review metrics. In other words, this study builds on the online consumer review literature by demonstrating how different collection methods of hospitality firms affect review metrics. It also provides recommendations regarding collection strategies to hospitality companies based on their needs. This study fills a literature gap by (1) systematically comparing different collection strategies used by hospitality firms, (2) investigating how collection methods affect key review metrics, and (3) comparing and contrasting the main methods of sentiment analysis in hotel reviews.

2 Background and Hypothesis Development

In the following section, I provide an overview of key literature on the use, collection and content contained within online reviews. Specifically, I explain why online word of mouth (OWOM) is so crucial to businesses, how online reviews may be biased, and how collection strategies can moderate collection bias. I draw upon the literature to establish a set of key hypotheses used to evaluate my research questions. Additionally, as the field of sentiment analysis is still evolving, I provide a brief overview, as it plays a key role in testing my hypotheses.

2.1 Online Word of Mouth

Word of mouth has been a topic of marketing research since the middle of the 20th century, but it has gained in exposure and importance since the emergence of the Internet (Trenz and Berger, 2013). OWOM has become an important source of publicity and provides crucial information during consumers' decision-making process (Li and Hitt, 2010). Among OWOM networks, the main form is online reviews. Online review data refer to"customer opinions (e.g., text reviews, numerical ratings, and personal information) left on online retailing and review platforms" (Tian et al., 2021).

Online reviews have become undeniably crucial for businesses. According to a report by Igniyte (2019), 93 percent of people say that online reviews impact their decisions, and 90 percent of consumers say that they take the time to read online reviews before visiting a business. Another review from BrightLocal (2019) found that 79 percent of consumers are as confident in online reviews as in personal recommendations from friends or family. Thus, the effect

that online reviews can have on businesses has become a prevalent topic in the marketing literature. Most articles have studied the effect of online reviews on sales, review helpfulness, or review manipulation (Trenz and Berger, 2013). Many studies have also investigated the impact of online reviews on consumer decision-making. Some studies have incorporated online review data to predict product sales and to create personalized recommendations.

Regarding the impact of online reviews on sales, Chevalier and Mayzlin (2006) analyzed how incremental change in the number of reviews affects book sales and showed that an improvement in book reviews leads to an increase in sales and that negative reviews have a stronger effect on sales than positive reviews. Rosario et al. (2016) conducted a meta-analysis across 96 studies covering 40 platforms and 26 product categories. Their study showed that on average, OWOM is positively related to sales but that the exact effects vary across platforms, products, and metrics. Blal and Sturman (2014) separated the effect of online reviews into volume and valence effects. These two main aspects of online reviews have different effects on hotels of different scales. Valence has a greater effect on higher-tier hotels' revPAR (revenue per available room), while volume has a greater effect on lower-tier hotels.

Exposure to online reviews also gives businesses more chances to be considered in consumers' minds. Online reviews can increase awareness of a product, and positive reviews can improve attitudes. This effect is stronger for lesser-known brands (Vermeulen and Seegers, 2009). Ghose and Ipeirotis (2011) explored multiple facets of online reviews, such as subjectivity levels, readability, and the extent of spelling errors, and found that these textual features matter for product sales and perceived usefulness. Ye et al. (2011) showed that a ten percent increase in an online review rating results in a more than 5 percent increase

in online booking intentions. In addition, online reviews can also affect product prices under dynamic pricing strategies like those used by hotels. Using transaction data from Travelocity, Anderson (2012) illustrated that an increase of one point in a hotel's online review score on a five-point scale can enable a hotel to increase its price by 11.2 percent and still maintain the same occupancy rate.

Based on the relationship between online reviews and sales, some researchers have used online review data to improve managers' decision-making process. Schneider and Gupta (2016) proposed a random projection approach to predict sales on Amazon.com using consumer reviews with an attributes-based regression model, and the predictive performance of this approach is strong.

2.2 Motivations of Posting Behavior

The motivation behind online posting behavior is rationalized by motivation theory, which classifies motivation as either intrinsic or extrinsic (Liang et al., 2017; Bilgram et al., 2008). Intrinsic motivation arises from inner feelings, such as the needs for uniqueness and attention (Khern-am nuai et al., 2018). Extrinsic motivation comes from external factors such as monetary and prosocial rewards. People who are extrinsically driven tend to exert the least effort to meet task requirements (Khern-am nuai et al., 2018). Askalidis et al. (2017) posited that the type of collection process can motivate different kinds of reviewers. Those who are self-motivated to leave a review can be regarded as intrinsically motivated, whereas those who are invited by retailers to leave a review can be classified as extrinsically motivated. According to motivation theory, intrinsically motivated reviewers should tend to give high-quality work. Using review data collected from four major online retailers, the authors tested the influence of the collection method (self-motivated versus retailer prompted) on review

ratings and length. The results show that self-motivated reviews tend to be longer and more negative (have a lower valence). It is possible that unsatisfied consumers are more self-motivated to leave reviews because of resentment or anger. In addition, negative consumers usually write detailed and lengthy reviews because negative moods are associated with the activation of detail-oriented systems (Bless et al., 1996; Bodenhausen et al., 1994; Schwarz, 1990).

2.3 Online Review Biases

Social exchange theory (SET) describes social interaction behavior as a consequence of a calculation of benefits and costs (Cook et al., 2013). Posting online reviews—sharing information with others online—is a type of social interaction behavior. Consumers are inclined to engage in this behavior when the perceived benefits outweigh the costs in their mind. Therefore, it is reasonable to infer that when the perceived motivation or costs—such as cognitive or time costs—change, consumers' information-sharing behavior may also change.

People with different experiences have different motivations for posting reviews. Due to these differences in motivation, online reviews may suffer from strong self-selection bias. Self-selection bias means that people who report their opinions on online review platforms are not fully representative of those who receive the associated services. A certain group of people might have a larger propensity to report their opinions.

Three forms of self-selection have been widely discussed in the online review literature: purchase self-selection (Hu and Pavlou, 2017; Kramer, 2007), intertemporal self-selection (Li and Hitt, 2008; Moe and Schweidel, 2012), and polarity self-selection (Hu and Pavlou, 2017). Purchase self-selection means that consumers who are satisfied are more likely to purchase a product. Customers

are likely to take online reviews into consideration while making a decision. Thus, businesses that already have positive reviews are likely to draw consumers' attention, and consumers may have a positive bias before they even visit the business and be more likely to write positive reviews after visiting the business (Hu et al., 2009; Moon et al., 2014). Intertemporal self-selection refers to different groups of consumers who self-select into leaving reviews at different stages of a product's life cycle. Li and Hitt (2008) demonstrated that earlier reviews tend to be positive and extreme due to differing reviewer profiles across the product life cycle (early versus late adopters). Polarity self-selection refers to selection bias caused by the different propensities to leave a review at different satisfaction levels. In general, there is a bimodal relationship between the satisfaction level and self-reported word-of-mouth intention, making the distribution of online review ratings J-shaped, meaning that people with extreme opinions are more likely to post reviews (Hu and Pavlou, 2017). In fact, Schoenmueller et al. (2020) tested the rating distribution on 25 major online review platforms and discovered that most platforms exhibit polarity and a positive imbalance, although the degree differs across platforms.

However, when the perceived posting cost changes, consumers behave differently. The above researchers also found that a mediator of self-selection bias is reviewers' familiarity with the online review platform. Schoenmueller et al. (2020) used the frequency with which a reviewer reviews on a platform as a proxy for polarity selection bias, and the results indicated that individuals who write few reviews tend to offer more polar reviews. When a consumer becomes more familiar with a platform and the posting process, their perceived posting cost decreases, leading to fewer polar reviews. Askalidis et al. (2017) found that email invitations can also reduce this selection bias in online reviews because

decreasing the perceived reporting cost can induce a new segment of customers to leave reviews. Therefore, it is reasonable to infer that for the group of reviewers whose reviews are prompted by email invitations, the motivation to post and the reporting propensity across all ratings are similar.

In this context, we are interested in the reasons behind different distributions of reviews from email-invited groups. The major difference across the email-prompted collection methods lies in the format that companies use to elicit reviews. The literature on satisfaction and psychometrics shows that scale modifications such as question framing, multi-item scales, and the number of questions can impact the review distribution (Danaher and Haddrell, 1996; Weijters et al., 2010; Moors et al., 2014).

Xiang et al. (2017) found different review sentiment and topic distributions among TripAdvisor, Expedia, and Yelp reviews. TripAdvisor and Yelp have much longer reviews with richer information than those on Expedia, where over 61 percent of reviews contain fewer than 25 words. The sentiment distribution was identified using lexicon-based sentiment analysis methods. Reviews on Expedia and TripAdvisor are skewed positively, whereas Yelp reviews are more polarized on both ends. The reasons behind these different distributions could relate to the different collection methods and scale modifications. However, the authors did not investigate which features of the platforms drive these differences. Kim and Hyun (2021) investigated one of the differences in the collection process—a social network interface system (SNIS) log-in feature—and tested how ease of log-in might influence review rating and length. The results show that allowing customers to log in using SNIS without creating a local TripAdvisor account induces lower ratings and shorter reviews. The SNIS feature can add convenience to the posting process, such that it attracts more people

into leaving reviews by decreasing the perceived posting cost, thereby inducing more consumers with nonextreme experiences to leave reviews. In other words, polarity decreases because of the increase in intermediate ratings. Based on the above literature, the next hypotheses is developed:

Hypothesis 1 (H1): *The lower the posting cost, the more people with lower intrinsic motivations are more likely to participate, and thus the polarity bias is reduced in ratings while the purchase self selection bias still maintains. Therefore, posting costs are inversely correlated with review scores.*

H1a. Platform rating scores: TA-SM <TA-GSS <TA-RE <Brand <Expedia

H1b. Collection attributes impact review scores.

Hypothesis 2 (H2): *The lower the posting cost, the more people with lower intrinsic motivations are more likely to participate, so there will be more less polar reviews. Posting costs are positively correlated with review score variances.*

H2a. Platform rating variances: Expedia <Brand <TA-RE <TA-SM <TA-GSS

H2b. Collection attributes impact review score variances.

2.4 Sentiment Analysis

Sentiment analysis is a well-known technique used to extract opinions or emotions from human-generated text. Sentiment analysis of hotel online reviews can be classified into two main categories: (1) machine learning approaches using techniques such as neural networks and (2) rule-based (lexicon-based) approaches, which rely on sentiment lexicons developed from the social media domain (Calheiros et al., 2017; Medhat et al., 2014).

Machine learning- (ML-) based approaches usually involve organizing a large amount of data under sentiment labels to train the ML model and applying the model to predict sentiments in future sentences. Typical ML approaches

used for sentiment analysis are naive Bayes, logistic regression, support vector machines, and deep learning (Chen et al., 2020; Chang et al., 2019). A very large dataset is needed to train an ML model. A good way to solve this problem is using BERT, created by Google in 2018 (Devlin et al., 2018). Unlike traditional techniques that analyze text from left-to-right or right-to-left, BERT is bidirectional, jointly conditioning on both left and right context in all layers. This process allows BERT to achieve high accuracy and incredible performance on small datasets, making it one of the most important and complete architectures for various natural language processing tasks (McGregor, 2020). Despite their disadvantages of high execution costs and tedious labeling work, ML-based methods demonstrate promising learning ability and high accuracy, so they are increasingly common in the industry.

Lexicon-based approaches usually utilize a certain dictionary indicating the polarity (positive or negative) of a list of words. Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of a text (Hatzivassiloglou and McKeown, 1997; Taboada et al., 2006). Next, sentiment is determined based on the appearance of the identified words in sentences. A combining function, such as averaging or summing, is then used to predict overall sentiment after adjustment of the sentence length.

The General Inquirer lexicon developed by Stone et al. (1962) has been widely used. It contains 1635 positive words and 2005 negative words. Drawing upon research conducted by Archak et al. (2011), Hu et al. (2012) extracted words from the General Inquirer lexicon and developed their own dictionary. Later, based on this approach, Calheiros et al. (2017) also developed specific lexicons for hotel online review sentiment classification.

Another widely used lexicon for online reviews is the NRC Emotion Lex-

icon (EmoLex). EmoLex is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiment valences (negative and positive). This work was manually compiled through crowd-sourcing.

Using the dictionary developed by Liu et al. (2010), Nielsen (2011), and Mohammad and Turney (2013), Mankad et al. (2016) utilized the EmoLex, which contains approximately 10000 labeled words, to predict the relationship between numerical ratings and sentiment. They found that reviews with stronger negative sentiment are correlated with lower ratings. In addition, negative sentiment has a stronger impact on rating scores than positive sentiment. Specifically, after controlling for other variables, the authors found that an additional negative word in a review can decrease a rating by 0.11 points, but an additional positive word in a review can increase a rating by only approximately 0.09 points.

Some researchers have also used the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon, as this lexicon performs exceptionally well in the social media domain (Hutto and Gilbert, 2014). The advantage of VADER is that the developers implemented heuristic rules that address punctuation, capitalization, adverbs, and contrasting conjunctions. However, VADER is specifically tuned to Twitter-like texts that are shorter than 280 characters and usually contain singular sentiment scores. Thus, it is not commonly used with online hotel reviews, as hotel reviews are usually longer and contain different sentiments within each review.

Lexicon-based methods have their own advantages. Unlike ML approaches, which require large data sets for training, lexicon-based approaches require no prior training. However, since the dictionary used is crucial, re-

searchers who use this method need to be discerning in regards to the dictionary domain. In addition, dictionaries usually require ongoing collection of new words and appearances online. Adding to or updating the dictionary requires domain expertise and tedious manual work. The accuracy rate of such methods is usually lower than that of ML-based methods because the former have no learning capabilities.

Geetha et al. (2017) explored the relationship between review rating and sentiment and found consistency between the two characteristics. Sentiment polarity accounts for significant variation in ratings across premium and budget hotels. Due to this high positive correlation between rating and sentiment, we assume that the effect of collection methods on ratings should be very similar to that on sentiment. Therefore, based on the above study, we explore several sentiment analysis methods for online reviews and test the following hypotheses:

Hypothesis 3 (H3): *The lower the posting cost, the more people with lower intrinsic motivations are more likely to participate, and thus the polarity bias is reduced in sentiment scores while the purchase self selection bias still maintains. Therefore, posting costs are inversely correlated with sentiment scores.*
H3a. Platform Sentiment Score: TA-SM <TA-GSS <TA-RE <Brand <Expedia
H3b. Collection attributes impact sentiment scores.

Hypothesis 4 (H4): *The lower the posting cost, the more people with lower intrinsic motivations are more likely to participate, so there will be more less polar reviews. Posting costs are positively correlated with sentiment variances. H4a. Platform Sentiment Variances: Expedia <Brand <TA-RE <TA-SM <TA-GSS*
H4b. Collection attributes impact sentiment variances.

3 Data, Methods and Results

The following section outlines the dataset used for analysis and the way that it was assembled, along with basic descriptive statistics of reviews. Two different modeling frameworks are outlined—one for analysis of numeric review outcomes (e.g., review scores, sentiment scores) and another to draw insights from qualitative review text (e.g., sentiment analysis).

3.1 Data Collection and Descriptive Statistics

To generate a representative sample of online hotel review collection methods, we selected four representative brands from the hotel industry: Marriott, Hilton, Choice, and Wyndham. These are the leading brands in the industry and have a fairly broad geographical distribution in the United States. By choosing these brands, we can rule out bias arising from deficiencies in the number or location of hotels. Additionally, Marriott and Choice use very different review collection methods from those of Hilton and Wyndham. The former two chains utilize the Brand and TA-RE methods, while the latter cooperate directly with TripAdvisor and collect reviews by TA-GSS. Due to these different strategies, we are able to obtain a representative sample of review collection methods.

After picking the four brands, we performed a simple random sample of the hotels in the major US geographical markets and obtained approximately 30 hotels at each scale level. The scale level is indicated by the Smith Travel Research (STR) hotel scale report. Web crawlers from Web Scraper were used to collect review information from the targeted hotels. Several types of data were collected, including rating, review content, collection method, and review time. The review data are limited to 2019 for the following reasons. First, review web-

sites and hotels were created in different years, and the full timeline of reviews cannot be captured. Second, many researchers have shown that the time trend can affect review patterns, and intertemporal self-selection bias can interfere with the effect of collection methods, which is what we want to measure. By limiting the review data to one year, we can reasonably rule out intertemporal self-selection bias in our models. In the end, sample data from 109947 reviews of 327 hotels were obtained.

Table 2 presents the hotel and review distribution for Marriott and Hilton. Marriott's total number of reviews on TripAdvisor and Expedia is much lower than Hilton's, accounting for nearly half the total for Hilton. Since Hilton hotels direct all their GSS reviews to TripAdvisor, it is reasonable to infer that Marriott would have a similar number of reviews on TripAdvisor if it were to also send all GSS reviews to TripAdvisor. By creating its own online review website, Marriott has formed a circular online community system and decreased its reliance on third-party websites to back up its reputation. However, as we discussed earlier, many consumers consider third-party review websites a more reliable information source. By having a dramatically greater number of reviews on TripAdvisor, Hilton might win public popularity through pure herd behavior. A similar pattern is shown in Table 3 for Choice and Wyndham. These two brands have a slightly different hotel profile from that of Marriott and Hilton, but Choice applies the same strategy as Marriott, and the strategy shows the same pattern as that observed for Marriott. It is worth considering which strategy is better for a brand seeking to boost its online reputation and popularity and whether it is worthwhile to cooperate with TripAdvisor to increase the number of published reviews.

Table 2: Sample Size: Hotels and Reviews by Scale for Marriott, Hilton

	Marriott			**Hilton**		
	UpperUp	Upscale	Mid	UpperUp	Upscale	Mid
Hotels	30	31	30	28	27	31
Reviews – TripAdvisor	2934	714	696	5477	3795	2250
Reviews – Brand.com	6599	2815	2802	n/a	n/a	n/a
Reviews – Expedia	8051	3191	3159	13598	8769	4383

Table 3: Sample Size: Hotels and Reviews by Scale for Choice, Wyndham

	Choice			**Wyndham**		
	Up	Mid	Econ	Up	Mid	Econ
Hotels	24	27	23	23	27	26
Reviews – TripAdvisor	1376	637	185	4654	820	494
Reviews – Brand.com	5647	2839	747	n/a	n/a	n/a
Reviews – Expedia	3680	3718	1436	8379	3871	2231

After classifying the reviews under the five collection methods, we can see the hotel and review distribution for each collection method, as shown in Table 4. Expedia has the most reviews because every hotel in our dataset uses Expedia to collect reviews. TA-RE reviews account for a small share of the total because only a few hotels have reviews on their official brand website while also cooperating with TripAdvisor and sending reviewers there. Only 23 Marriott and Choice hotels choose to perform both brand collection and TA-RE collection.

Table 4: Hotel and Review Distribution by Collection Method

Method	**Brand**	**Expedia**	**TA-GSS**	**TA-SM**	**TA-RE**
Number of Hotels	165	327	160	323	23
Number of Reviews	21449	64466	9249	14088	695

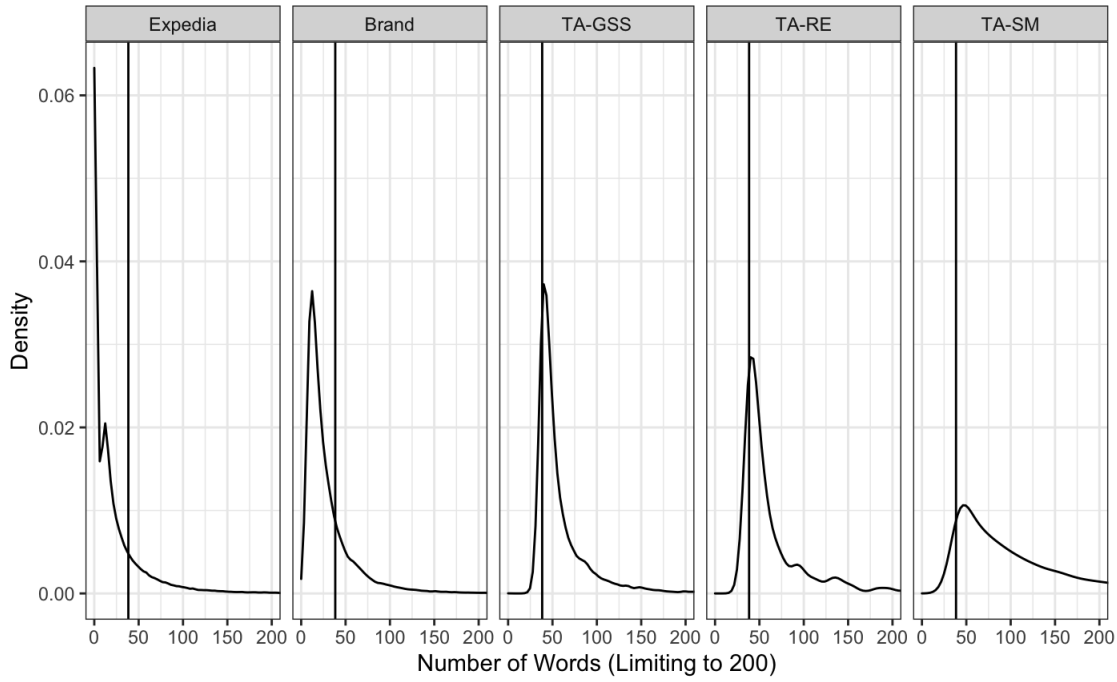
The main review statistics are presented in Table 5. The ratings in our dataset are positively skewed across all collection methods. Partially aligning with Hypothesis 1, TA-SM shows a relatively low average rating in comparison with those of other email-solicited methods.TA-SM reviews are the longest. TA-RE has the highest mean rating.

Table 5: Review Descriptive Statistics by Collection Method

Method	Mean Rating	Median Rating	SD Rating	Mean Length	Median Length	SD Length
Expedia	4.10	4	1.12	19.81	9	33.74
Brand	4.24	5	1.18	30.45	20	31.88
TA-GSS	3.89	5	1.38	60.17	47	39.13
TA-RE	4.41	5	1.03	66.36	49	47.19
TA-SM	3.90	4	1.29	119.84	86	106.39
Method	# Reviews	% of 1	% of 2	% of 3	% of 4	% of 5
Expedia	64466	0.05	0.06	0.11	0.30	0.48
Brand	21449	0.06	0.06	0.08	0.20	0.61
TA-GSS	9249	0.10	0.09	0.12	0.18	0.50
TA-RE	695	0.04	0.03	0.07	0.19	0.66
TA-SM	14088	0.09	0.08	0.14	0.25	0.45

We can take a closer look at the review length feature in Figure 7. The x-axis represents the number of words in each review, and we set the upper bound to 200 because every method has the same long tail trend after 200 words. The y-axis represents the density of a certain review length. The vertical line represents the average number of words (38) in our whole dataset. The review length for Expedia is substantially skewed toward the shorter end, with 43.6 percent of reviews being empty. In fact, the vast majority (83.5 percent) of reviews on Expedia have fewer than 38 words, while this is true of only 76 percent of the Brand reviews, 13 percent of TA-GSS review, 15 percent of TA-RE reviews, and 5 percent of TA-SM reviews.

Figure 7: Length Distribution by Collection Method



The rating distribution across all collection methods for all hotels is plotted

in Figure 8. Simply looking at the plot, we can expect considerable variance in the average rating across collection methods within each hotel and across the hotels.

Figure 8: Rating Distribution by Hotel

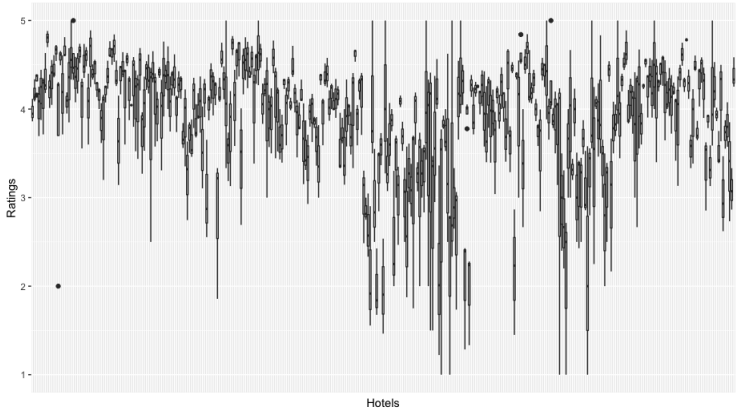


Table 6 displays the collection methods used by each major brand. Marriott and Choice collect reviews through GSS and post the reviews to their brand websites. Hilton and Wyndham collect reviews through GSS and post the reviews to Tripadvisor. A few hotels from Marriott and Choice choose to additionally utilize the Review Express method from TripAdvisor. Table 6 shows the mean rating by platform, brand, and scale. Table 7 shows the mean length by platform, brand, and scale. Every brand is using TripAdvisor and Expedia while only two of them are using the brand platform. Across all brands, Expedia generates higher ratings with shorter length whereas TA generates lower ratings with longer review. Nearly all reviews from the Expedia method have a mean length of less than 25 words, which makes the Expedia method unfavorable for use in the later sentiment analysis section because of the limited information that it provides.

Table 6: Mean Rating by Collection Platform, Brand, and Scale

Collection		Brand	Expedia	TripAdvisor
Marriott	**All**	**4.2**	**4.3**	**4.2**
	Upper Up	4.2	4.3	4.2
	Upscale	4.3	4.4	4.1
	Mid	4.4	4.4	4.6
Choice	**All**	**4.2**	**4.2**	**4.3**
	Upscale	4.4	4.5	4.5
	Mid	4.1	4.2	4.2
	Econ	3.5	3.5	3.2
Hilton	**All**		**4.1**	**3.9**
	Upper Up		4.1	3.8
	Upscale		4.1	3.8
	Mid		4.4	4.1
Wyndham	**All**		**3.7**	**3.7**
	Upscale		4.0	3.8
	Mid		3.2	3.0
	Econ		3.6	3.6

Table 7: Mean Length by Collection Platform, Brand, and Scale

Collection		Brand	Expedia	TripAdvisor
Marriott	**All**	**33**	**17**	**98.0**
	Upper Up	34	17	103.6
	Upscale	31	16	98.7
	Mid	31	16	75.8
Choice	**All**	**28**	**20**	**77.4**
	Upscale	28	20	78.7
	Mid	27	19	71.0
	Econ	30	23	93.2
Hilton	**All**		**19**	**92.9**
	Upper Up		19	96.5
	Upscale		19	87.9
	Mid		21	93.0
Wyndham	**All**		**24**	**89.6**
	Upscale		23	91.9
	Mid		26	86.2
	Econ		22	76.4

Before we conduct the sentiment analysis, null and non-English reviews need to be removed from the dataset. Null reviews are reviews with ratings but no review content. Using the Langdetect package in Python 1936, non-English reviews were detected. Both null and non-English reviews were removed. In the end, 28062 null and 1936 non-English reviews were removed, yielding a dataset of 79946 reviews. Table 8 displays the number of reviews kept and removed for each collection method. As the table indicates, except for Expedia, the collection methods are impacted very little by the exclusion of reviews that contain no text responses. Table 9 summarizes the percent of reviews with text (Nonnull) and without text (Null) across the different numeric review score categories (1-5).

Table 8: Reviews Kept and Removed for Each Collection Method

	Brand	Expedia	TA-GSS	TA-RE	TA-SM
Reviews Kept	21315	34602	9248	695	14086
Reviews Removed	134	29864	1	0	2

As Table 9 indicates that the distributions of nonnull versus null reviews across scores are not similar, we formalize these differences with a chi-square test. We conduct the chi-square test on the rating distributions for Expedia reviews to see if the removal affects the overall distribution of potential sentiment. We do not conduct the tests for other collection methods because the number of reviews removed for those methods is rather small, and we assume the removal does not significantly affect the distribution of sentiment. The chi-square test for the rating distributions of reviews kept and removed for Expedia shows a significant difference between the distributions ($p<2.2e-16$). Furthermore, Table 9 indicates that null reviews have a higher percentage of 4 and 5 ratings. After removing all these null reviews for the sentiment score model, we have

fewer positive reviews. Based on the high positive correlation between rating and sentiment, the sentiment of Expedia in reality should be more positive, but since it is not possible to obtain the sentiment for null reviews, we are unable to determine the real sentiment distribution for all Expedia reviews.

Table 9: Rating Distribution of Nonnull vs. Null Expedia Reviews

Numeric Rating	Nonnull Reviews	Null Reviews
1	0.059	0.033
2	0.079	0.037
3	0.135	0.087
4	0.270	0.327
5	0.455	0.514

3.2 Text Mining and Sentiment Analysis

As introduced in the literature, sentiment analysis has been widely used for online hotel reviews and is still rapidly evolving. Both lexicon and ML methods are explored within this study. For lexicon-based methods, two lexicons are chosen: VADER and EmoLex. For ML methods, the up-to-date BERT model is utilized and trained with three different approaches. The first training label is provided by public IMDB review data with 2 polarities. The second training label is provided by public Rotten Tomatoes review data with 5 sentiment levels. The third training label is sampled from our dataset, and we use the rating score as the label for sentiment levels. The sample dataset with 79946 nonnull reviews is used, with 10000 of them randomly selected to serve as the training dataset. To maintain the consistency of the data used to analyze the collection effect, the remaining 69946 reviews are used as the final dataset for the full sentiment analysis.

3.2.1 Preprocessing for Lexicon-Based Sentiment Analysis

Table 10 shows the preprocessing of the review texts used for lexicon-based modeling. All English reviews collected under the five collection methods are preprocessed through the following steps: lower-case, tokenize, remove numbers, remove stop words, and lemmatize. Lower-casing, number removal and stop word removal are used to eliminate unhelpful parts of the data. Stop words are words that do not contribute to the meaning of a text. The NLTK package in Python provides 179 English stop words. Tokenizing transforms each individual sentence into a lexicon-readable format. In our case, it is used to split sentences into words. Lemmatization groups all kinds of forms of the same word to align with the form included in the lexicons. For example, after lemmatization, "staffs" is transformed to "staff".

The two lexicons are selected based on former research (Liu et al., 2010; Mankad et al., 2016; Hutto and Gilbert, 2014). One is the VADER lexicon, which contains 11000 words. The other lexicon is EmoLex, which contains 14181 words. How the preprocessed words are coded in the VADER lexicon and EmoLex is provided in Table 11. It is worth noting that neither lexicon assigns sentiment values to most nouns or verbs. It is mostly adjectives that carry sentiment values.

Table 10: Preprocessing Steps

Lowercase	Tokenize	Punctuation, Number, Stopwords	Lemmatize
no staffs was available to make breakfast on 12/26 morning, Coffee service available, but was cold and weak coffee. The uncomfortable room couch was also dirty.	no		
	staffs	staffs	staff
	was		
	available	available	available
	to		
	make	make	make
	breakfast	breakfast	breakfast
	on		
	12/26		
	morning,	morning	morning
	coffee	coffee	coffee
	service	service	service
	available	available	available
	but		
	was		
	cold	cold	cold
	and		
	weak	weak	weak
	coffee.	coffee	coffee
	the		
	uncomfortable	uncomfortable	uncomfortable
	room	room	room
	couch	couch	couch
	was		
	also	also	also
	dirty	dirty	dirty

Table 11: Word Sentiment Examples from Lexicons

	Vader		EmoLex	
	Polarity	Intensity	Positive	Negative
staff				
available				
make				
breakfast			1	0
morning				
coffee				
service				
cold			0	1
weak	-1.9	0.7	0	1
uncomfortable	-1.6	0.5	0	1
room				
couch				
also				
dirty	-1.9	0.80	0	1

3.2.2 Training for Machine Learning-Based Sentiment Analysis

ML-based sentiment analysis methods include both training and predicting processes. In the training process, the model learns to associate a specific input (i.e., texts) with corresponding tags (e.g., positive, neutral, negative) based on the data samples used for training. The training data contain the texts (reviews) and the tags (polarity or emotions). Thus, a correctly labeled training dataset is crucial for ML-based methods. These training data usually require expensive class labeling. First, two publicly available datasets with training labels for online reviews are collected. The IMDB dataset has two polarity tags (positive vs. negative). The Rotten Tomatoes (RT) dataset has five classification labels, similar to the rating levels in our data.

The IMDB review dataset from the popular movie rating service was col-

lected and prepared by Andrew L. Maas (Maas et al., 2011). It contains 25000 reviews for training and 25000 reviews for testing. All these reviews are labeled, indicating whether the reviews are positive or negative. It is a binary classification. A sample of the training data is provided in Table 23.

Table 12: IMDB Training Data Sample

Review	Sentiment
Story of a man who has unnatural feelings for a pig. Starts out with a opening scene that is a terrific example of absurd comedy. A formal orchestra audience is turned into an insane, violent mob by the crazy chantings of it's singers. Unfortunately it stays absurd the WHOLE time with no general narrative eventually making it just too off putting. Even those from the era should be turned off. The cryptic dialogue would make Shakespeare seem easy to a third grader. On a technical level it's better than you might think with some good cinematography by future great Vilmos Zsigmond. Future stars Sally Kirkland and Frederic Forrest can be seen briefly.	3
Robert DeNiro plays the most unbelievably intelligent illiterate of all time. This movie is so wasteful of talent, it is truly disgusting. The script is unbelievable. The dialog is unbelievable. Jane Fonda's character is a caricature of herself, and not a funny one. The movie moves at a snail's pace, is photographed in an ill-advised manner, and is insufferably preachy. It also plugs in every cliche in the book. Swoozie Kurtz is excellent in a supporting role, but so what? Equally annoying is this new IMDB rule of requiring ten lines for every review. When a movie is this worthless, it doesn't require ten lines of text to let other readers know that it is a waste of time and tape. Avoid this movie.	1
BromIll High is nothing short of brilliant. Expertly scripted and perfectly delivered, this searing parody of a students and teachers at a South London Public School leaves you literally rolling with laughter. It's vulgar, provocative, witty and sharp. The characters are a superbly caricatured cross section of British society (or to be more accurate, of any society). Following the escapades of Keisha, Latrina and Natella, our three "protagonists" for want of a better term, the show doesn't shy away from parodying every imaginable subject. Political correctness flies out the window in every episode. If you enjoy shows that aren't afraid to poke fun of every taboo subject imaginable, then BromIll High will not disappoint!	9

The Rotten Tomatoes dataset was collected by Pang and Lee (2005). It con-

tains 156060 phrases from the reviews on Rotten Tomatoes. Each phrase is labeled on a 5-level sentiment scale: negative, somewhat negative, neutral, somewhat positive, and positive. A sample of the training data is provided in Table 13.

Table 13: Rotten Tomatoes Training Data Sample

Phrase	Sentiment
A series of escapades demonstrating the adage that what is good for the goose is also good for the gander, some of which occasionally amuses but none of which amounts to much of a story.	1
A series of escapades demonstrating the adage that what is good for the goose	2
A series	2
this Oscar-nominated documentary takes you there.	3
this Oscar-nominated documentary	4
Oscar-nominated documentary	4
Oscar-nominated	4
... the whole thing succeeded only in making me groggy.	1
the whole thing succeeded only in making me groggy.	1
succeeded only in making me groggy.	3

Since these two public datasets are both for movie reviews, they have limitations in predicting sentiment in hotel reviews. In an ideal world, sentiment labels could be manually coded by different coders based on our hotel dataset. However, this is not feasible for our process, so we deal with the missing hotel training label problem by sacrificing objectivity. A total of 10000 reviews are randomly selected to work as the training dataset, with the label indicators being review ratings from 1 to 5.

3.3 Methods and Results

To measure the effect of collection methods on review rating and sentiment, we employ a linear regression model because our main goal is to obtain a simple yet interpretable output. In section 3.1, Figure 8 displays the rating distribution for each hotel across all collection methods. The significant variation in ratings across hotels makes it necessary to control for hotel characteristics. Since specific hotel-level data, such as hotel size and age, are limited, to control for hotel heterogeneity in our model, we use a mixed effect model. The review data are clustered because they are repeated data for each individual hotel over time. A mixed effect model can obtain cluster-specific (hotel) effects in addition to the standard coefficients in a regression model. This choice offers the option to capture cluster-specific effects while borrowing the strengths of fixed effect variables. The fixed effects estimation controls for time-invariant differences across hotels, and the random effects estimation controls for unobservable factors that may also influence review metrics. As we want to measure how the effect of the collection method differs within each hotel, we treat the collection method as a fixed effect variable. The hotel variable is coded as a random variable to control for hotel-level heterogeneity. Under other circumstances, we could include the brand and scale variables, but considering that brands actually indicate which collection method hotels use, the model would be overspecified with the inclusion of both brand and collection method. In addition, by specifying the hotel, the scale is also specified, so including these variables in the model would be redundant. To measure the effect directly and interpret the model clearly, we decide to keep the model simple while controlling for the necessary interference factors.

3.3.1 Collection Effect on Ratings

To test Hypothesis 1a regarding the collection effect on ratings, we first conduct an ANOVA for collection method effects on ratings; the result (F=215.9, $p <$ 2.2e-16) shows a significant difference in the effects of collection methods on ratings. With ratings taken as the dependent variable to fit the mixed effect model discussed above, the regression results follow in Table 14. The reference collection method is Expedia. Compared to Expedia, Brand, TA-GSS, and TA-SM have significantly different effects. The effect of TA-RE does not show a significant difference from that of Expedia.

Table 14: Model: Collection Method Effect on Ratings

Random effects:		
Groups	Variance	Std Dev
Hotel	0.2381	0.488
Residual	1.2299	1.109
Number of obs: 109947 groups: hotel, 327		

Fixed effects:	Estimate	p-value
(Intercept)	4.09E+00	<2e-16
collectionBrand	-8.69E-02	<2e-16
collectionTA-GSS	-2.16E-01	<2e-16
collectionTA-RE	3.97E-03	0.93
collectionTA-SM	-2.80E-01	<2e-16

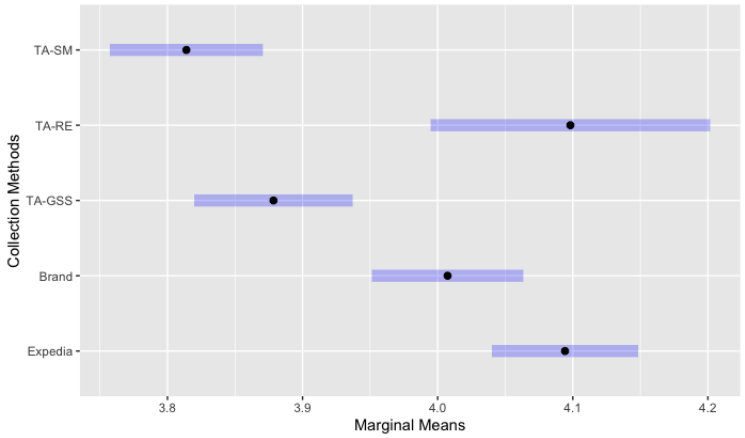
To further investigate how the effects of collection methods differ, we compare the effects of collection methods pairwise, as shown in Table 15. Specifically, the pairwise comparison table shows that the Expedia method effect is not significantly different from that of TA-RE, and the Brand method effect is not significantly different from that of TA-RE due to the wide confidence intervals of the latter.

Table 15: Pairwise Collection Method Marginal Mean Comparison

Pairwise Contrast	Estimate	p-value
Expedia – Brand	0.08694	<.0001
Expedia – (TA-GSS)	0.2158	<.0001
Expedia – (TA-RE)	-0.00397	1
Expedia – (TA-SM)	0.28026	<.0001
Brand – (TA-GSS)	0.12886	<.0001
Brand – (TA-RE)	-0.09091	0.268
Brand – (TA-SM)	0.19332	<.0001
(TA-GSS) – (TA-RE)	-0.21977	<.0001
(TA-GSS) – (TA-SM)	0.06445	0.0003
(TA-RE) - (TA-SM)	0.28423	<.0001

Figure 9 shows the marginal effect of collection methods on ratings after inclusion of controls for hotel heterogeneity. The rating scores induced by the collection methods are ordered as follows: TA-SM < TA-GSS < Brand < Expedia < TA-RE.

Figure 9: Marginal Means by Collection Method



The findings show that self-motivated reviews have lower ratings than email-prompted reviews. This confirms the prediction of motivation theory that people with extreme and negative experiences are more likely to be self-motivated to post reviews. However, this result is different from the expectations in the general order of Hypothesis 1a. We cannot see a clear increasing trend in the ratings as posting costs of the collection method decrease.

To test Hypothesis 1b regarding how each posting cost attribute affect review ratings, we applied the mixed effect model and got the following regression output in Table 16. The regression results prove that having posting costs will reduce the ratings for most posting cost attributes. In our model, after controlling the effect of hotels, having login method can decrease the rating by 0.13 and having survey can decrease the rating by 0.22. Having email invitation (which can reduce the posting cost) can increase the rating by 0.28. The only attribute different from expectation is detail (i.e. require the minimum words). The result here shows that having some detail requirements can actually increase the rating by 0.13.

Table 16: Posting Cost Attribute Effects

	Coefficients	P-value
(Intercept)	3.81	<2e-16 ***
login	-0.13	4.20e-15 ***
survey	-0.22	3.21e-06 ***
email	0.28	7.34e-10 ***
detail	0.13	0.00703 **

To test Hypothesis 2a regarding collection effect on rating variances, we obtained the variances of the five collection methods as table 17. We further did a pairwise variance significance test as shown in table 18. The results indicate the order of rating variances is: TA-RE <Expedia <Brand <TA-SM <TA-GSS.

Table 17: Comparison for Rating Variances

	Brand	Expedia	TA-GSS	TA-SM	TA-RE
Variance	1.39	1.27	1.91	1.67	1.05

Table 18: Pairwise Comparison for Rating Variances Significance

	Brand	Expedia	TA-GSS	TA-RE
Expedia	<2e-16	-	-	-
TA-GSS	<2e-16	<2e-16	-	-
TA-RE	1.50E-06	0.0011	<2e-16	-
TA-SM	<2e-16	<2e-16	4.60E-13	1.30E-14

To test Hypothesis 2b regarding posting cost attribute effect on rating variances, we obtained the variances of ratings for each cost attribute and did an F-test to check the significance. The results in table 19 indicate that this hypothesis is correct, which means lower posting costs (no login, no survey, with email invitation, no detail) generate lower variances in ratings.

Table 19: Comparison for Variances by Cost Attribute

Cost	login	survey	email	detail
0 (No)	1.30	1.34	1.67	1.27
1 (Yes)	1.75	1.57	1.36	1.61
F-Test	***	***	***	***

3.3.2 Collection Effect on Sentiments

After we apply the five sentiment analysis methods, the sentiment score data are obtained. All sentiment scores are scaled to the range of -1 to 1, where -1 means extremely negative, 0 means neutral, and 1 means extremely positive. To directly compare the scores, we use two measures: classification accuracy metrics and the sentiment score distribution.

By approaching review sentiment analysis as a classification task, i.e., a simple test of whether a review is positive or negative, we can use classification accuracy metrics to compare the model performance. To check the accuracy rate, we create a proxy for the sentiment indicator from our own dataset. Assuming reviews with rating 1 to be negative and reviews with rating 5 to be positive, we can obtain the classification accuracy performance of all five methods. PR-AUC and the F1-score are used to evaluate the accuracy. PR-AUC stands for the area under the precision and recall curve, and a score of 1.0 represents a model with perfect accuracy. The F1-score represents a balance between precision and recall, and the closer the score is to 1 the better (Brownlee, 2020). Table 20 shows the PR-AUC and F1-score for each sentiment analysis method. In general, the ML methods provide better classification accuracy.

Table 20: PR-AUC and F1-Score for Sentiment Methods

Method	PR-AUC	F1-Score
VADER	0.92	0.93
EmoLex	0.87	0.90
BERT IMDB	0.97	0.94
BERT RT	0.98	0.86
BERT Rating	0.98	0.97

To develop further into the sentiment analysis for each review to examine not only the sentiment polarity classification but also the sentiment degree, we create an overall sentiment boxplot for clear comparison. To clearly display how each sentiment analysis method performs for negative, neutral, and positive reviews, we manually group reviews with ratings 1 and 2 as negative, reviews with rating 3 as neutral, and reviews with ratings 4 and 5 as positive.

Figure 10 shows the sentiment distribution using the lexicon-based method with the VADER lexicon. We can see that the mean sentiment for negative and

neutral reviews is much higher than it is supposed to be. This indicates that this method performs poorly at detecting negative sentiment and may classify overall sentiment scores in the dataset as higher than they are in reality. This result may also be because VADER is tuned to shorter texts, so it may not be able to evaluate all of the words in an analyzed review. In many longer reviews, people tend to talk about the things that they think the hotels do well first, followed by a "but", where they may start to talk about negative things.

Figure 10: Sentiment Distribution by VADER

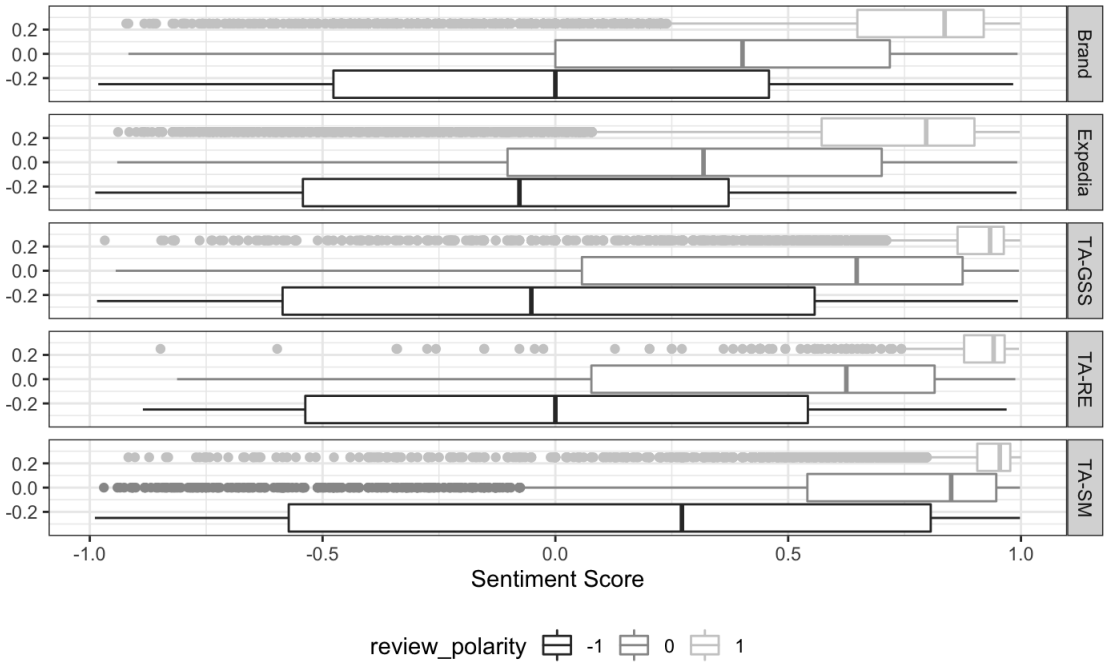


Figure 11 shows the sentiment distribution using the lexicon-based method with EmoLex. Interestingly, all three methods are grouped near the neutral sentiment, which indicates that EmoLex may not assign a very clear sentiment degree but just a simple polarity.

Figure 11: Sentiment Distribution by EmoLex

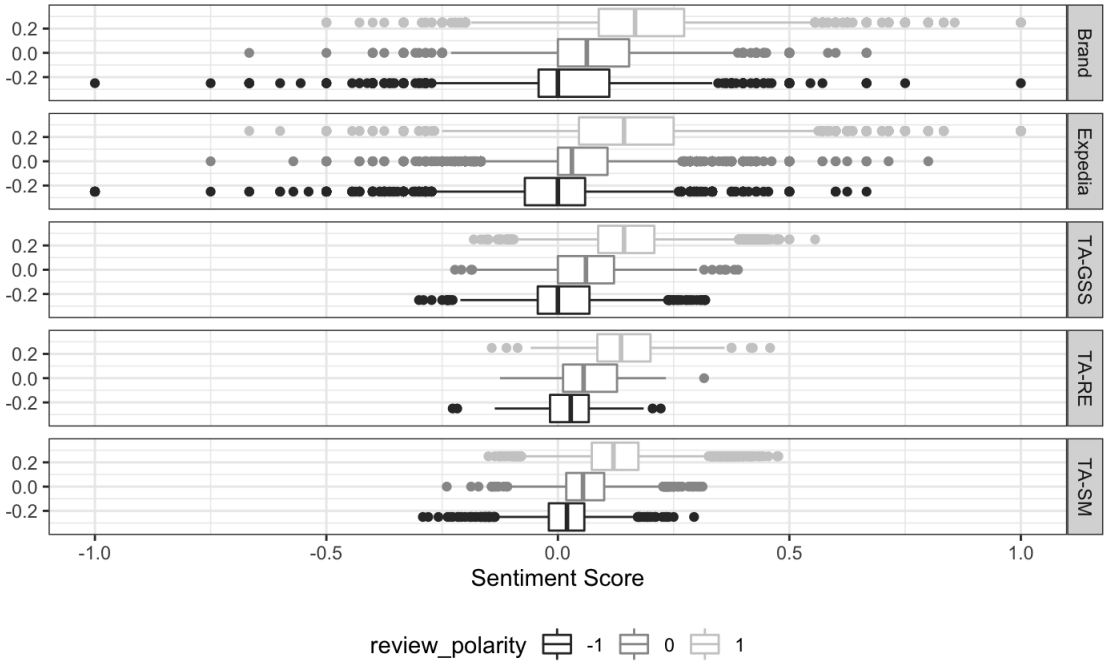


Figure 12 shows the sentiment distribution using the ML-based method with IMDB labels. We can see a similar pattern in the sentiment scores to that of the rating distribution figure above. However, all the sentiment scores for neutral reviews skew negatively, which means that this method may overstate negative sentiment relative to reality. In addition, the means of positive and negative sentiment reviews tend toward the extremes. This could be because this IMDB dataset is a binary classification dataset that assigns sentiment scores more extreme values, meaning it can correctly represent only the sentiment polarity, not the sentiment degree.

Figure 13 shows the sentiment distribution using the ML-based method with Rotten Tomatoes labels. This method also shows that neutral reviews have relatively negative sentiment. This could be because of a model deficiency or because of inconsistency between ratings and sentiment; that is, people might

give neutral ratings but actually express negative sentiments in their reviews. In addition, the model performs fairly well at classifying positive and negative reviews with reasonable mean values and ranges.

Figure 14 shows the sentiment distribution using the ML-based method with rating labels. This model is trained directly using the rating labels in our dataset, so it should have the most similar pattern to that of the rating distributions. However, it assigns positive sentiment to most neutral rating reviews. Again, this could be due to a model deficiency or to inconsistency between ratings and sentiment. It is also possible that the reviews used for training are not very clean and that the training sample is too small to actually train the model well.

Figure 12: Sentiment Distribution by BERT Using IMDB Labels

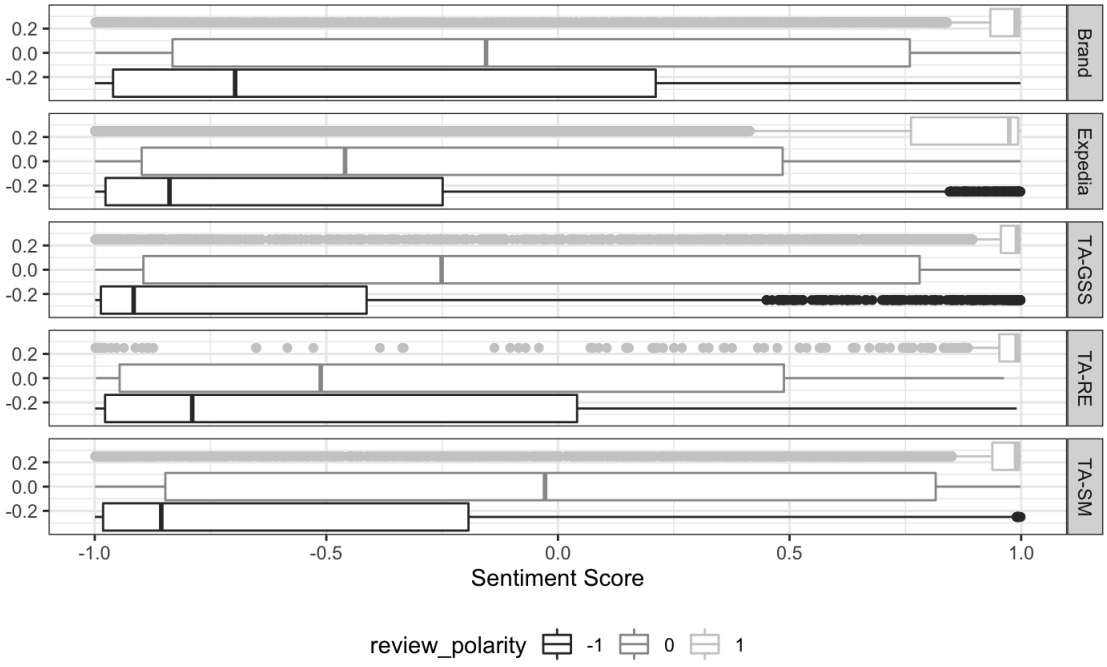


Figure 13: Sentiment Distribution by BERT Using Rotten Tomato Data Labels

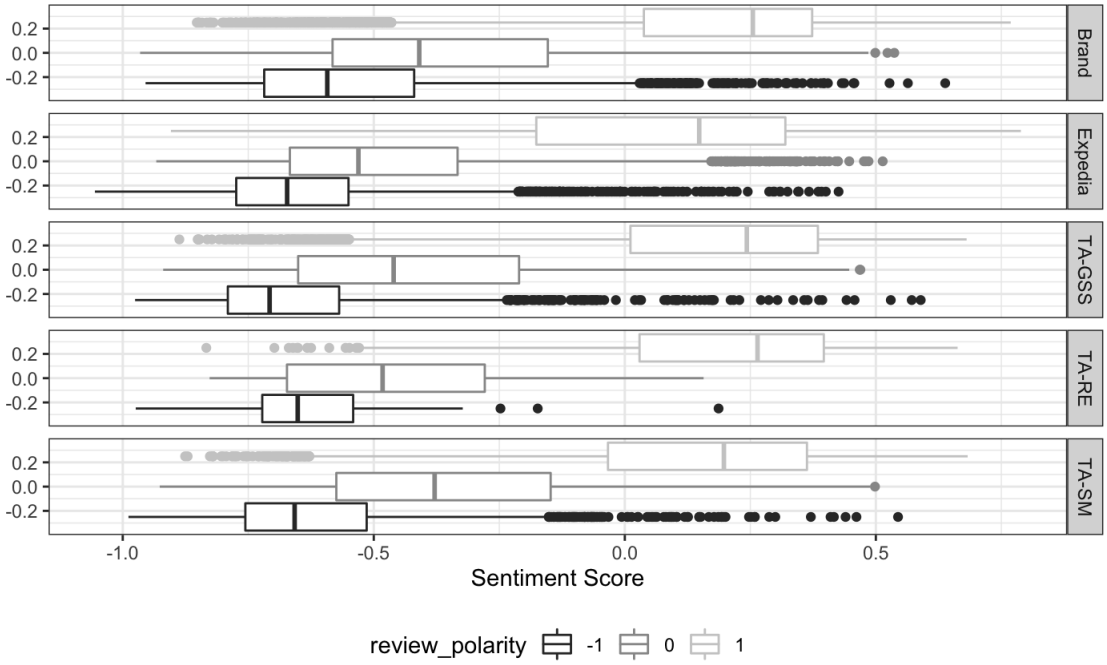
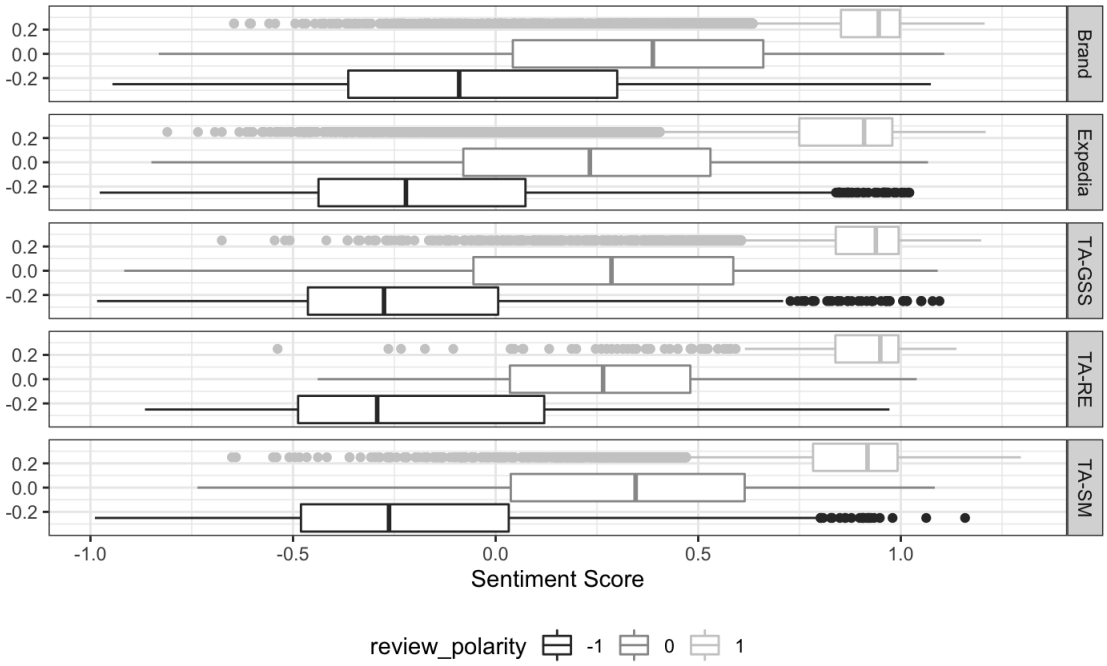


Figure 14: Sentiment Distribution by BERT Using Rating Labels



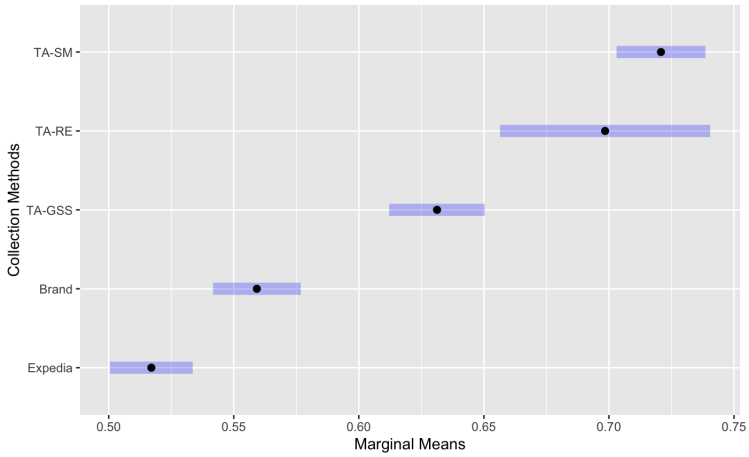
To test Hypothesis 3a, same logic is applied as in rating section but with sentiment scores as the dependent variable. The sentiment scores induced by each collection method is expected this order: TA-SM <TA-GSS <TA-RE <Brand <Expedia. Using sentiment scores from VADER method as the dependent variable, the regression outputs is obtained (Table 21). Reference collection method is Expedia. To further investigate how the effects of collection methods differ, a pairwise comparison between collection methods is obtained (Table 21), which shows that only TA-RE method is not significantly different from TA-SM.

Table 21: Collection Effect on Vader Scores

Fixed effects:	Estimate	p-value
(Intercept)	5.17E-01	<2e-16
collectionBrand	4.22E-02	3.90E-16
collectionTA-GSS	1.14E-01	<2e-16
collectionTA-RE	1.81E-01	<2e-16
collectionTA-SM	2.04E-01	<2e-16
Pairwise Contrast	Estimate	p-value
Expedia – Brand	-0.0422	<.0001
Expedia – (TA-GSS)	-0.1142	<.0001
Expedia – (TA-RE)	-0.1814	<.0001
Expedia – (TA-SM)	-0.2038	<.0001
Brand – (TA-GSS)	-0.072	<.0001
Brand – (TA-RE)	-0.1392	<.0001
Brand – (TA-SM)	-0.1616	<.0001
(TA-GSS) – (TA-RE)	-0.0672	0.0111
(TA-GSS) – (TA-SM)	-0.0895	<.0001
(TA-RE) – (TA-SM)	-0.0223	0.8065

Figure 15 shows the marginal effect of collection methods on sentiment after inclusion of controls for hotel heterogeneity. The rating scores induced by the collection methods are ordered as follows: Expedia $<$ Brand $<$ TA-GSS $<$ TA-RE/TA-SM. As we discussed earlier, the VADER method classifies many neutral and negative sentiments as positive, so the result quite possibly does not reflect the true sentiment scores. Since lexicon-based methods give scores based on words, the TA method could be favored because of more positive words. It may be the case that an unsatisfied reviewer on TripAdvisor uses many positive words and mentions just one negative thing, such as horrible cleanliness, and still gives a rating of 1.

Figure 15: Marginal Means by Collection Method on VADER Sentiment



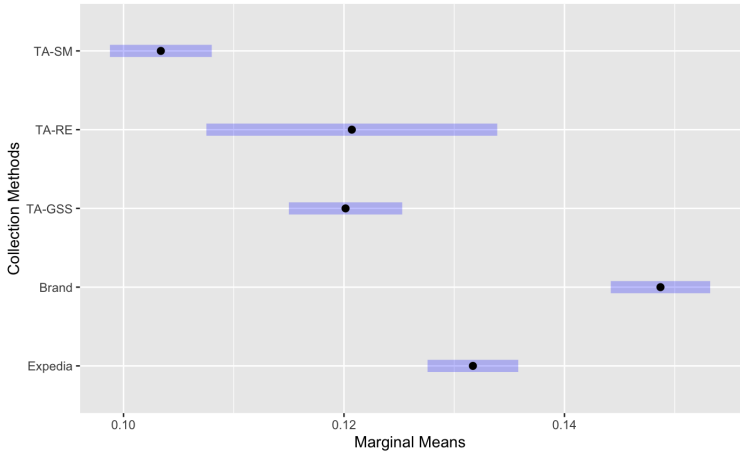
Using sentiment scores from the EmoLex method as the dependent variable, we have the regression outputs in Table 22. The reference collection method is Expedia. Compared to Expedia, all other collection methods except TA-RE have significantly different effects. To further investigate how the effects of collection methods are different, we show a pairwise comparison of the effects of collection methods in Table 22. Specifically, the pairwise comparison table shows that the TA-RE method effect is not significantly different from those of Expedia, TA-SM or TA-GSS.

Table 22: Collection Effect on EmoLex Scores

Fixed effects:	Estimate	p-value
(Intercept)	1.32E-01	<2e-16
collectionBrand	1.70E-02	<2e-16
collectionTA-GSS	-1.16E-02	3.22E-09
collectionTA-RE	-1.10E-02	0.0923
collectionTA-SM	-2.83E-02	<2e-16
Pairwise Contrast	Estimate	p-value
Expedia – Brand	-0.017017	<.0001
Expedia – (TA-GSS)	0.011553	<.0001
Expedia – (TA-RE)	0.010984	0.4443
Expedia – (TA-SM)	0.028309	<.0001
Brand – (TA-GSS)	0.02857	<.0001
Brand – (TA-RE)	0.028	0.0002
Brand – (TA-SM)	0.045325	<.0001
(TA-GSS) – (TA-RE)	-0.000569	1
(TA-GSS) – (TA-SM)	0.016756	<.0001
(TA-RE) – (TA-SM)	0.017325	0.0651

Figure 16 shows the marginal effect of collection methods on sentiment after we control for hotel heterogeneity. The rating scores induced by the collection methods are ordered as follows: TA-SM <TA-GSS/TA-RE < Expedia < Brand. However, as we discussed earlier, the EmoLex method does not have a clear classification, so the result quite possibly does not capture the true degree of the sentiment scores.

Figure 16: Marginal Means by Collection Method on EmoLex Sentiment



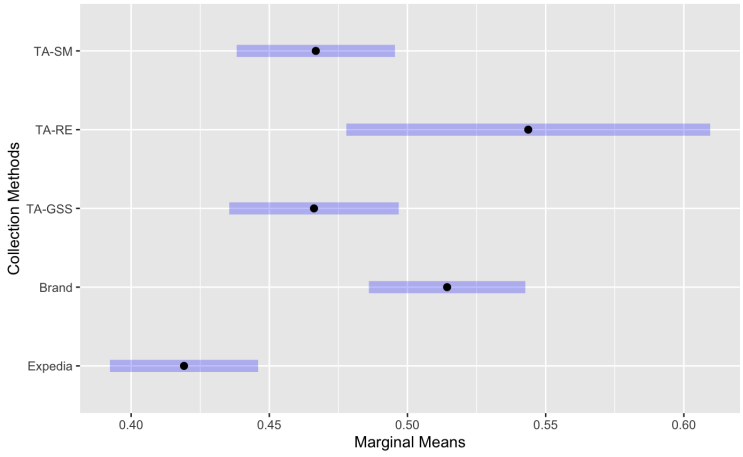
Using sentiment scores from the BERT IMDB method as the dependent variable, we have the regression outputs in Table 23. The reference collection method is Expedia. Compared to Expedia, all other collection methods have significantly different effects. To further investigate how the effects of collection methods differ, we show a pairwise comparison in Table 23. Specifically, the pairwise comparison table shows that the effect of the TA-RE method is not significantly different from that of TA-SM or Brand. TA-GSS does not have significantly different effect from that of TA-SM.

Table 23: Collection Effect on BERT IMDB Scores

Fixed effects:	Estimate	p-value
(Intercept)	4.19E-01	<2e-16
collectionBrand	9.52E-02	<2e-16
collectionTA-GSS	4.70E-02	5.35E-07
collectionTA-RE	1.25E-01	6.92E-05
collectionTA-SM	4.77E-02	8.75E-10
Pairwise Contrast	**Estimate**	**p-value**
Expedia – Brand	-0.095194	<.0001
Expedia – (TA-GSS)	-0.047024	<.0001
Expedia – (TA-RE)	-0.124552	0.0007
Expedia – (TA-SM)	-0.047681	<.0001
Brand – (TA-GSS)	0.04817	0.0004
Brand – (TA-RE)	-0.029358	0.8795
Brand – (TA-SM)	0.047513	<.0001
(TA-GSS) – (TA-RE)	-0.077528	0.1185
(TA-GSS) – (TA-SM)	-0.000657	1
(TA-RE) – (TA-SM)	0.076871	0.1067

Figure 17 shows the marginal effect order of collection methods: Expedia $<$ TA-GSS/TA-SM $<$ Brand/TA-RE.

Figure 17: Marginal Means by Collection Method on BERT IMDB Sentiment



Using sentiment scores from the BERT RT method as the dependent variable, we obtain the regression outputs in Table 24. The reference collection method is Expedia. Compared to Expedia, all other collection methods have significantly different effects. To further investigate how the effects of collection methods vary, we present a pairwise comparison in Table 24. Specifically, the pairwise comparison table shows that the effect of the TA-RE method is not significantly different from those of Brand or TA-GSS. TA-GSS does not have a significantly different effect from that of TA-SM.

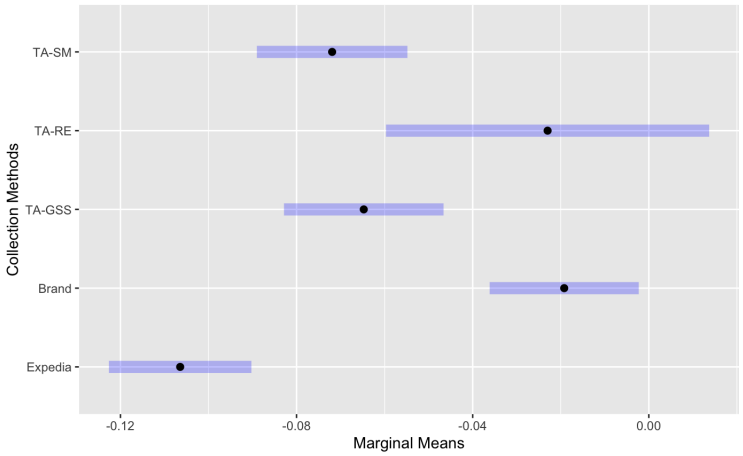
Table 24: Collection Effect on BERT RT Scores

Fixed effects:	Estimate	p-value
(Intercept)	-1.06E-01	<2e-16
collectionBrand	8.72E-02	<2e-16
collectionTA-GSS	4.17E-02	5.08E-16
collectionTA-RE	8.35E-02	1.14E-06
collectionTA-SM	3.45E-02	5.48E-16
Pairwise Contrast	Estimate	p-value
Expedia – Brand	-0.08722	<.0001
Expedia – (TA-GSS)	-0.0417	<.0001
Expedia – (TA-RE)	-0.08345	<.0001
Expedia – (TA-SM)	-0.03451	<.0001
Brand – (TA-GSS)	0.04552	<.0001
Brand – (TA-RE)	0.00377	0.9995
Brand – (TA-SM)	0.0527	<.0001
(TA-GSS) – (TA-RE)	-0.04175	0.1303
(TA-GSS) – (TA-SM)	0.00718	0.7328
(TA-RE) – (TA-SM)	0.04893	0.0379

Figure 18 shows the marginal effect of collection methods on sentiment after inclusion of controls for hotel heterogeneity. The rating scores induced by the collection methods are ordered as follows: Expedia $<$ TA-SM/TA-GSS $<$

TA-RE/Brand. In the case of Expedia, it has relatively low sentiment. On the one hand, this could be because of a model deficiency since many Expedia observations were removed from the dataset. On the other hand, it is also possible that people who write reviews at Expedia tend to write negative reviews.

Figure 18: Marginal Means by Collection Method on BERT RT Sentiment



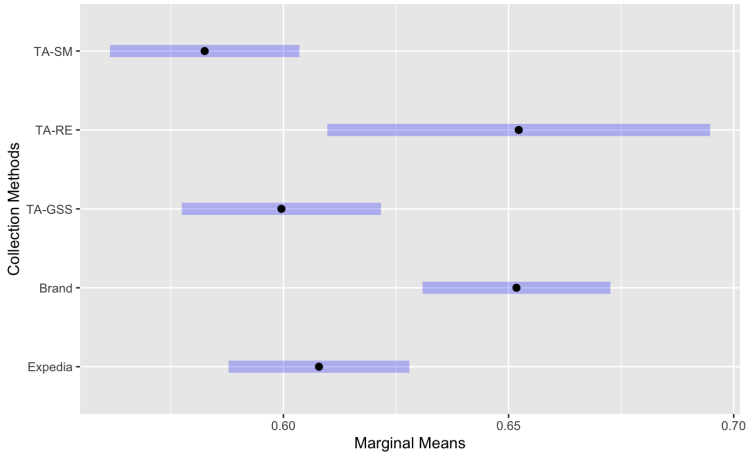
Using sentiment scores from the BERT rating method as the dependent variable, we obtain the regression outputs in Table 25. The reference collection method is Expedia. Compared to Expedia, all other collection methods except TA-GSS have significantly different effects. To further investigate how the effects of collection methods differ, we present a pairwise comparison in Table 29. Specifically, the pairwise comparison table shows that the Expedia method does not have a significantly different from those of TA-GSS or TA-RE. The TA-RE method effect is not significantly different from that of Brand. The TA-GSS effect is not significantly different from those of TA-RE or TA-SM.

Table 25: Collection Effect on BERT Rating Sentiment

Fixed effects:	Estimate	p-value
(Intercept)	6.08E-01	<2e-16
collectionBrand	4.39E-02	<2e-16
collectionTA-GSS	-8.35E-03	0.154
collectionTA-RE	4.44E-02	0.023
collectionTA-SM	-2.54E-02	1.62E-07
Pairwise Contrast	Estimate	p-value
Expedia – Brand	-0.043862	<.0001
Expedia – (TA-GSS)	0.008345	0.6107
Expedia – (TA-RE)	-0.044371	0.1533
Expedia – (TA-SM)	0.025391	<.0001
Brand – (TA-GSS)	0.052207	<.0001
Brand – (TA-RE)	-0.000508	1
Brand – (TA-SM)	0.069253	<.0001
(TA-GSS) – (TA-RE)	-0.052715	0.0696
(TA-GSS) – (TA-SM)	0.017046	0.0764
(TA-RE) – (TA-SM)	0.069761	0.0036

Figure 19 shows the marginal effect order of collection methods on sentiment scores: TA-SM <TA-GSS/Expedia <Brand/TA-RE.

Figure 19: Marginal Means by Collection Method on BERT Rating Sentiment



To test Hypothesis 3b regarding the effect of posting cost attribute on review sentiment scores, we apply the mixed effect model and use the 5 sentiment

scores as dependent variables individually. The regression output is shown in Table 26 for Lexicon-based sentiment scores and Table 27 for ML-based sentiment scores. H3b is partially correct. For most (4 of 5) sentiment analysis methods, lower posting costs induce higher sentiment scores with one exception – the detail attribute. We expect having detail as increasing cost to decrease sentiment score, but in fact having detail can increase the sentiment score, which aligns with our findings for rating scores as well.

Table 26: Regression Output for Lexicon-based Scores

	Vader		EmoLex	
	Estimate	P-value	Estimate	P-value
(Intercept)	0.54	***	1.14E-01	***
login	0.07	***	-2.86E-02	***
survey	-0.07	**	-5.69E-04	
email	-2.23E-02		1.73E-02	**
detail	1.09E-01	***	1.76E-02	*

Table 27: Regression Output for ML-based Scores

	BERT-IMDB		BERT-RT		BERT-Rating	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
(Intercept)	3.42E-01	***	-1.55E-01	***	5.38E-01	***
login	-4.82E-02	***	-4.55E-02	***	-5.22E-02	***
survey	-7.75E-02	*	-4.18E-02	*	-5.27E-02	**
email	7.69E-02	*	4.89E-02	**	6.98E-02	***
detail	1.73E-01	***	1.29E-01	***	9.66E-02	***

In general, the expected and tested result from Hypothesis 3 is shown in table 28.

Table 28: Result Summary of Hypothesis 3

Method	Expected Effect	Lexicon-Vader	Lexicon-EmoLex	BERT-IMDB	BERT-RT	BERT-Rating
H3a	S <G<R<B<E	E<B<G<R/S	S<G/R<E<B	E<G/S<B/R	E<G/S<B/R	S<G/E<B/R
H3b - login	-	+	-	-	-	-
H3b - survey	-	-	-	-	-	-
H3b - email	+	-	+	+	+	+
H3b - detail	-	+	+	+	+	+

To test Hypothesis 4, we follow the same process for Hypothesis 3 and get the following result in Table 29. H4a is not correct and H4b is partially correct. Most sentiment methods (4 of 5) suggest that having login will increase the sentiment variances as we expected. The consistent result out of the 5 sentiment methods is survey and detail attributes, which contradict with our expectations. We expect having survey and detail (increase the cost) to induce higher variances, however, the sentiment outputs suggest that having survey and detail can actually lower the variances. The effect of emails on sentiment variances differs a lot from method to method, so it is hard to draw a conclusion.

Table 29: Result Summary of Hypothesis 4

Method	Expected Effect	Lexicon-Vader	Lexicon-EmoLex	BERT-IMDB	BERT-RT	BERT-Rating
H4a	E<B<R<S<G	R<B<S<E<G	S<R<G<B<E	R<B<S<E<G	R/B<S<E<G	R/B<E<S<G
H4b - login	+	+	-	+	+	+
H4b - survey	+	-	-	-	-	-
H4b - email	-	+	+	-	+	-
H4b - detail	+	-	-	-	-	-

4 Discussion and Limitations

Among the different firm strategies for review collection, we explored five different kinds of collection methods utilized by hospitality companies and further identified the effect of these collection methods and posting cost attributes on review metrics such as rating and sentiment. We find some consistencies between our hypotheses and the real data distribution. First, self-motivated consumers tend to give lower ratings and longer reviews. The general pattern between posting costs and review rating and length roughly aligns with our hypotheses. It is worth noting that businesses need to increase posting costs to some extent so that consumers provide some detailed information and put more thought into the review. However, companies also need to find a balance between motivations and costs so that consumers are motivated to consider their reviews carefully but do not lose patience during the process. How to strike this balance properly is a question worth considering in future research. Second, it is consistent that the lower the posting cost, the lower the variances between ratings, which means the polarity selection bias is reduced. To generate more representative and objective opinions from all consumers, companies should consider utilizing the posting motivation and cost theory to design effective forms. Third, directing GSS reviews to TripAdvisor or using TA's Review Express can help companies get higher ratings with more information on TA platform. However, for the design of hotels' GSS form, instead of doing both private survey and TripAdvisor review process in GSS, hotels should send GSS that contain either the private survey for brands or the TripAdvisor review invitations to different subgroups of consumers because this is more cost efficient for consumers.

With respect to the sentiment analysis process, there are several limitations. First, the relationship between rating and sentiment is still debated, but to sim-

plify the analysis, we assume that the two variables are positively correlated, regardless of whether the reviews are negative, neutral, or positive. However, in reality, certain ratings may not positively correlate with sentiment. Second, in the current public dictionaries, many hotel-specific words are not included, and the sentiment level may diverge because dictionaries are usually domain specific. Ideally, we would use a hotel-based sentiment dictionary for lexicon-based sentiment analysis. Third, due to time and financial constraints, we are unable to label the online hotel reviews manually from different objective coders and instead utilize public movie review labels as substitutes in the ML-based sentiment analysis. Fourth, if we had more detailed hotel-level data, we could better control for potential confounding variables in our model.

Nevertheless, this study helps us gain a better understanding of the biases in online reviews, moderating factors to control these biases, and the corresponding effects on review metrics of these moderators.

References

Anderson, C. (2012). The impact of social media on lodging performance.

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.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management science*, 57(8):1485–1509.

Askalidis, G., Kim, S. J., and Malthouse, E. C. (2017). Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:23–30.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817.

Bilgram, V., Brem, A., and Voigt, K.-I. (2008). User-centric innovations in new product development—systematic identification of lead users harnessing interactive and collaborative online-tools. *International journal of innovation management*, 12(03):419–458.

Blal, I. and Sturman, M. C. (2014). The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales. *Cornell Hospitality Quarterly*, 55(4):365–375.

Bless, H., Clore, G. L., Schwarz, N., Golisano, V., Rabe, C., and Wölk, M. (1996). Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of personality and social psychology*, 71(4):665.

Bodenhausen, G. V., Kramer, G. P., and Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of personality and social psychology*, 66(4):621.

BrightLocal (2019). Local consumer review survey 2019.

Brownlee, J. (2020). Roc curves and precision-recall curves for imbalanced classification.

Calheiros, A. C., Moro, S., and Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 26(7):675–693.

Chang, Y.-C., Ku, C.-H., and Chen, C.-H. (2019). Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. *International Journal of Information Management*, 48:263–279.

Chen, W., Xu, Z., Zheng, X., Yu, Q., and Luo, Y. (2020). Research on sentiment classification of online travel review text. *Applied Sciences*, 10(15):5275.

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.

Cook, K. S., Cheshire, C., Rice, E. R., and Nakagawa, S. (2013). Social exchange theory. *Handbook of social psychology*, pages 61–88.

Danaher, P. J. and Haddrell, V. (1996). A comparison of question scales used for measuring customer satisfaction. *International Journal of Service Industry Management*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training

of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DMR (2021). Tripadvisor statistics, user counts, and facts.

Geetha, M., Singha, P., and Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management*, 61:43–54.

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Godes, D. and Mayzlin, D. (2009). Firm-created word-of-mouth communication: Evidence from a field test. *Marketing science*, 28(4):721–739.

Guo, Y., Barnes, S. J., and Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59:467–483.

Han, S. and Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, 61(2):142–153.

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, pages 174–181.

Hu, H.-H., Kandampully, J., and Juwaheer, T. D. (2009). Relationships and impacts of service quality, perceived value, customer satisfaction, and image: an empirical study. *The service industries journal*, 29(2):111–125.

Hu, N., Bose, I., Koh, N. S., and Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 52(3):674–684.

Hu, N. and Pavlou, P. A. (2017). ON SELF-SELECTION BIASES IN ONLINE PRODUCT REVIEWS. *MIS Quarterly*, 41(2):1–17.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Igniyte (2019). 30 essential online review facts and stats.

Kankanhalli, A., Tan, B. C., and Wei, K.-K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS quarterly*, pages 113–143.

Khern-am nuai, W., Kannan, K., and Ghasemkhani, H. (2018). Extrinsic versus intrinsic rewards for contributing reviews in an online platform. *Information Systems Research*, 29(4):871–892.

Kim, J. M. and Hyun, S. (2021). Differences in online reviews caused by distribution channels. *Tourism Management*, 83(March 2019):104230.

Klein, N., Marinescu, I., Chamberlain, A., and Smart, M. (2018). Online reviews are biased. here's how to fix them. *Harvard Business Review*.

Kramer, M. A. (2007). Self-selection bias in reputation systems. In *IFIP International Conference on Trust Management*, pages 255–268. Springer.

Kumar, N. and Benbasat, I. (2006). Research note: the influence of recommenda-

tions and consumer reviews on evaluations of websites. *Information Systems Research*, 17(4):425–439.

Lee, M. K., Cheung, C. M., Lim, K. H., and Sia, C. L. (2006). Understanding customer knowledge sharing in web-based discussion boards: An exploratory study. *Internet Research*.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Li, X. and Hitt, L. M. (2010). Price effects in online product reviews: An analytical model and empirical analysis. *MIS quarterly*, pages 809–831.

Liang, S., Zhang, Z., Zhang, Z., Law, R., and Sun, W. (2017). Consumer motivation in providing high-quality information: building toward a novel design for travel guide websites. *Asia Pacific Journal of Tourism Research*, 22(6):693–707.

Liang, T.-P., Liu, C.-C., and Wu, C.-H. (2008). Can social exchange theory explain individual knowledge-sharing behavior? a meta-analysis. *ICIS 2008 proceedings*, page 171.

Liu, B. et al. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Liu, X., Schuckert, M., and Law, R. (2016). Online incentive hierarchies, review extremity, and review quality: Empirical evidence from the hotel sector. *Journal of Travel & Tourism Marketing*, 33(3):279–292.

Liu, X., Schuckert, M., and Law, R. (2018). Utilitarianism and knowledge growth during status seeking: Evidence from text mining of online reviews. *Tourism Management*, 66:38–46.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Mankad, S., Han, H. Goh, J., and Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. *Service Science*, 8(2):124–138.

McGregor, M. (2020). Google bert nlp machine learning tutorial.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Miguéns, J., Baggio, R., and Costa, C. (2008). Social media and tourism destinations: Tripadvisor case study. *Advances in tourism research*, 26(28):1–6.

Min Kim, J., Han, J., and Jun, M. (2020). Differences in mobile and nonmobile reviews: the role of perceived costs in review-posting. *International Journal of Electronic Commerce*, 24(4):450–473.

Moe, W. W. and Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Moon, S., Park, Y., and Kim, Y. S. (2014). The impact of text product reviews on sales. *European Journal of Marketing*.

Moors, G., Kieruj, N. D., and Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, 44(1):369–399.

Mudambi, S. M. and Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Nielsen, F. . (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Papathanassis, A. and Knolle, F. (2011). Exploring the adoption and processing of online holiday reviews: A grounded theory approach. *Tourism Management*, 32(2):215–224.

Research, P. (2019). Four key developments keeping the u.s. ota market exciting.

Rosario, A. B., Sotgiu, F., De Valck, K., and Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, 53(3):297–318.

Schneider, M. J. and Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243–256.

Schoenmueller, V., Netzer, O., and Stahl, F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research*, 57(5):853–877.

Schwarz, N. (1990). *Feelings as information: Informational and motivational functions of affective states.* The Guilford Press.

Statista (2021). Gross bookings of expedia group, inc. worldwide from 2005 to 2020.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., and Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484.

Taboada, M., Anthony, C., and Voll, K. D. (2006). Methods for creating semantic orientation dictionaries. In *LREC*, pages 427–432.

Tian, G., Lu, L., and McIntosh, C. (2021). What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data. *Food Quality and Preference*, 88(November 2019):104060.

Trenz, M. and Berger, B. (2013). Analyzing online customer reviews - An interdisciplinary literature review and research agenda. *ECIS 2013 - Proceedings of the 21st European Conference on Information Systems*.

Vermeulen, I. E. and Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1):123–127.

Wasko, M. M. and Faraj, S. (2005). Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pages 35–57.

Weijters, B., Cabooter, E., and Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247.

Wu, T.-Y. and Lin, C. A. (2017). Predicting the effects of ewom and online brand messaging: Source trust, bandwagon effect and innovation adoption factors. *Telematics and Informatics*, 34(2):470–480.

Xiang, Z., Du, Q., Ma, Y., and Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58:51–65.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., and Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 44:120–130.

Xiong, G. and Bharadwaj, S. (2014). Prerelease buzz evolution patterns and new product performance. *Marketing Science*, 33(3):401–421.

Ye, Q., Law, R., Gu, B., and Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human behavior*, 27(2):634–639.