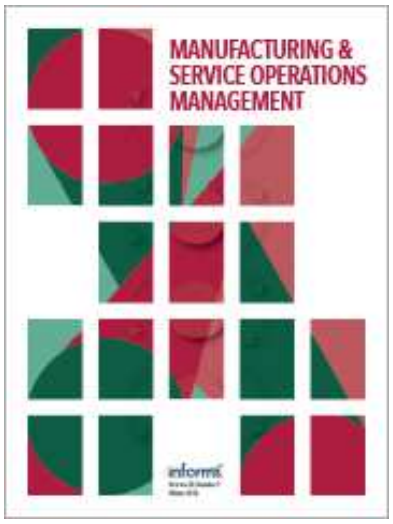

Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Service Quality Using Text Mining: Measurement and Consequences

Jorge Mejia, Shawn Mankad, Anandasivam Gopal

To cite this article:
Jorge Mejia, Shawn Mankad, Anandasivam Gopal (2021) Service Quality Using Text Mining: Measurement and Consequences. Manufacturing & Service Operations Management 23(6):1354-1372. https://doi.org/10.1287/msom.2020.0883

http://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 23, No. 6, November–December 2021, pp. 1354–1372
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Service Quality Using Text Mining: Measurement and Consequences

Jorge Mejia,[a] Shawn Mankad,[b] Anandasivam Gopal[c]

[a] Kelley School of Business, Indiana University, Bloomington, Indiana 47405; [b] Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853; [c] Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742

Contact: jmmejia@iu.edu, https://orcid.org/0000-0002-4479-1209 (JM); spm263@cornell.edu, https://orcid.org/0000-0001-7945-8556 (SM); agopal@rhsmith.umd.edu, https://orcid.org/0000-0001-9270-5961 (AG)

Received: February 1, 2018
Revised: January 3, 2019; October 4, 2019
Accepted: January 12, 2020
Published Online in Articles in Advance: August 4, 2020

https://doi.org/10.1287/msom.2020.0883

Abstract. ***Problem description***: Measuring quality in the service industry remains a challenge. Existing methodologies are often costly and unscalable. Furthermore, understanding how elements of service quality contribute to the performance of service providers continues to be a concern in the service industry. In this paper, we address these challenges in the restaurant sector, a vital component of the service industry. ***Academic/practical relevance***: Our work provides a scalable methodology for measuring the quality of service providers using the vast amount of text in social media. The quality metrics proposed are associated with economic outcomes for restaurants and can help predict future restaurant performance. ***Methodology***: We use text present in online reviews on Yelp.com to identify and extract service dimensions using nonnegative matrix factorization for a large set of restaurants located in a major city in the United States. We subsequently validate these service dimensions as proxies for service quality using external data sources and a series of laboratory experiments. Finally, we use econometrics to test the relationship between these dimensions and restaurant survival as additional validation. ***Results***: We find that our proposed service quality dimensions are scalable, match industry standards, and are correctly identified by subjects in a controlled setting. Furthermore, we show that specific service dimensions are significantly correlated with the survival of merchants, even after controlling for competition and other factors. ***Managerial implications***: This work has implications for the strategic use of text analytics in the context of service operations, where an increasingly large text corpus is available. We discuss the benefits of this work for service providers and platforms, such as Yelp and OpenTable.

Funding: This work was supported by National Science Foundation [Grant 1633158].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/msom.2020.0883.

Keywords: service quality • word of mouth • business survival • econometrics • text analysis • machine learning

1. Introduction

In a comprehensive review of service operations, Chase and Apte (2007) argue that establishing a clear relationship between service quality and performance is an important component of service operations management. Service quality has been a central theme in this field, as either the focal response variable or a key independent variable. Ideally, high service quality should lead to customer loyalty, increased revenue, positive word of mouth, and higher stock market valuation (Zeithaml et al. 1996, Ramdas et al. 2013). Yet, finding appropriately scalable and available measures of service quality on providers for use over time has proven difficult (Voss et al. 2008, Rosenzweig et al. 2011). Two main strategies are used to overcome this challenge: surveys, such as ServQual (Parasuraman et al. 1988, Roth and Jackson 1995), measure service quality directly; alternatively, indirect indicators such as *wait times* (Allon et al. 2011b, Ibrahim and Whitt 2011b, Ang et al. 2015, Batt and Terwiesch 2015), queue length (Veeraraghavan and Debo 2008), and service errors (Soteriou and Zenios 1999, Xu and Chan 2016) can be used as proxies for service quality.

Unfortunately, both approaches have well-documented shortcomings. With surveys, it is challenging to construct measures of service quality that are generalizable across providers (Roth and Jackson 1995) or even across time for the same provider, especially as the service mix changes. Moreover, surveys are time-consuming and costly, even when performed correctly (Cronin and Taylor 1992). Meanwhile, indirect measures are idiosyncratic to the particular provider and not easy to use across providers (Roth and Menor 2003). Moreover, scalability remains a problem, especially across providers and over time. Therefore, the problem of measuring service quality, as a precursor to

understanding the link to performance, continues to remain relevant.

In this paper, we argue that online reviews represent a third potential strategy to effectively measure service quality, in addition to surveys and indirect measures. Platforms like Yelp and TripAdvisor provide forums for customers to contribute direct and unfiltered feedback to service providers. In the digital platforms literature, online reviews are viewed as important sources of customer feedback since they are free, are easily available, and represent collective opinions (Dellarocas 2003). Positive online reviews have also been associated with performance outcomes, such as online and offline sales, and have been shown to influence customers' choice of service providers (Chevalier and Mayzlin 2006, Duan et al. 2008, Ye et al. 2009, Cui et al. 2020). Prior work has typically used the star rating and the number of reviews to represent the traffic or popularity of the provider (Dellarocas and Narayan 2006). However, this ignores the text consumers write, which represents the majority of review content (Cao et al. 2011). We focus on the text as a source of service quality information that may provide ubiquity, scale, and relevance to service providers.

Although some existing research has explored text from online reviews (Archak et al. 2011, Ghose and Ipeirotis 2011, Mankad et al. 2016), we use recent advances in natural language processing to summarize the review text in terms of latent themes pertaining to service quality. We focus on restaurants within a large North American city. Restaurants are an ideal context for this work since online reviews are well-established and valued by service providers (Lu et al. 2013). We use a data set of online reviews from Yelp (Yelp 2017), including over 130,000 reviews comprising approximately 50,000 pages of text for over 2,400 restaurants over a nine-year period.

We evaluate multiple text mining methodologies to identify and extract service quality information from online review text. We use nonnegative matrix factorization (NMF), a well-established text mining methodology (Lee and Seung 1999, Xu et al. 2003, O'Callaghan et al. 2015), to identify semantic components that pertain to the service quality of restaurants. In robustness checks described in online Appendix B, we evaluate several alternative methodologies including aspect-based semantic analysis (ABSA), which extracts semantic service quality components as well as sentiment (Brody and Elhadad 2010, Wang et al. 2010). Our primary NMF analysis identified five components reflecting aspects of a restaurant's service: *overall quality*, *wait times*, *food quality*, *responsiveness*, and *atmosphere*. Using these components, we can quantify restaurants' service quality over time and compare restaurants in a multidimensional and scalable manner. Critically, we also validate the extracted semantic components by testing their correspondence with other observed characteristics and performance outcomes through three stages.

First, we ask, do the identified semantic components provide a reasonable classification of the text in online reviews? In other words, from a behavioral perspective, do human subjects respond similarly to reviews that are characteristic of the semantic dimensions we identify? Such a test is important to establish that the semantic dimensions are comparable to human subjects' interpretation of reviews in a controlled environment. In order to establish this concordance, we conduct a series of experiments on Amazon Mechanical Turk (MTurk) where we test if subjects respond to reviews selected to reflect the semantic themes in terms of purchase intention and *overall quality*, thereby "reverse engineering" our text mining results and providing a source of external validation. We find strong support for these tests: semantic dimensions loading highly on specific service quality dimensions are classified as such by subjects at statistically significant levels.

Second, we examine the semantic themes' external validity by linking them to established measures of service quality provided by an industry standard, Zagat. Zagat provides annual quality scores for restaurants along three dimensions: food, decor, and service, thus providing a comparable "expert" rating of service quality. We find substantial agreement in dimensions and yearly scores. In summary, these validation tests show strong support for the use of our semantic themes as reflections of restaurants' service quality.

Our final validation test addresses the link between the services the restaurant provides in a competitive marketplace and its performance. Existing theoretical frameworks (Roth and Jackson 1995, Heskett et al. 2008) contend that service quality, reflecting superior service design, should lead to better economic performance (Goldstein et al. 2002). In our validation test, we use semantic features to predict an important economic outcome—*survival* (Rosenzweig et al. 2011). If the provider's financial viability is linked to service quality (Voss et al. 2008), lower service quality should significantly increase the odds of business failure. Combining a panel data set of restaurant closures with the semantic service quality scores, we estimate econometric models of firm survival and find that the semantic themes significantly increase the explanatory power of such models, relative to models with only numerical review data and control variables (Parsa et al. 2005, 2011). We find that the *overall quality* and *wait times* semantic themes are significantly associated with survival, consistent with prior survey-based work (Allon et al. 2011b). In terms of prediction,

the inclusion of the semantic themes in the model improves longitudinal classification accuracy by 70%. Although we do not make causal arguments about the effects of online reviews on survival, this analysis illustrates that semantic quality scores can provide valuable insights to restaurateurs, real estate agents, and other financial stakeholders.

Our work makes several contributions to the service operations literature. First, as an increasing number of service encounters involve "experiences" (Pine and Gilmore 1999, Heineke and Davis 2007), unstructured feedback, such as online review text, is increasingly relevant. Second, the work we describe here bridges an existing gap between service quality, online reviews, and firm survival (Voss et al. 2008, Rosenzweig et al. 2011, Ramdas et al. 2013). Finally, we are able to assess dimensions that are hard to measure at scale and across the competitive marketplace, such as the efficiency and effectiveness of service encounters. In contrast to previous work that has used primary data to measure wait times in contexts where such data are available, such as in call centers and airlines (Kim and Whitt 2014, Ang et al. 2015), we use review text to generate indicators of wait times and other quality measures related to survival and purchase intention.

Our findings have direct implications for the service industry. Using review-based semantic themes, restaurants can identify service elements needing improvement and benchmark their results to others in the market. Potential investors and landlords, lacking access to detailed financial records, can gauge the health of service providers. Platforms like Yelp can provide insights to businesses and consumers. This approach is also applicable in other service contexts where text-based reviews are available, such as hotels and retail.

2. Background and Theory

2.1. Measuring Quality in Service Operations

Understanding the drivers and consequences of service quality has been a significant part of the research agenda in service operations (Heineke and Davis 2007). However, the question of how to measure service quality and customer satisfaction effectively on an ongoing basis and at scale has remained a significant challenge (Rust et al. 1995). To this end, a variety of methods to measure service quality have been used in the literature. Early work addressing service quality extended metrics of quality from the manufacturing context using numerical indicators that reflected a production process approach (Wyckoff 2001), such as service "defects" (Hall and Porteus 2000), dissatisfied customer communication (Soteriou and Chase 2000), and "service errors" (Soteriou and Zenios 1999). However, since services require a different approach emphasizing "service experience" and "engagement" (Pine and Gilmore 1999, Heineke and Davis 2007), quality metrics from manufacturing do not address these nuances effectively.

More recently, scholars have focused on estimating customer *wait times* as a potential proxy for service quality (Armony et al. 2008, Ibrahim and Whitt 2011b, Ibrahim and L'Ecuyer 2012). Many service contexts treat *wait times* as a measure of service quality and advertise them as such (Allon et al. 2011b). However, much of the work in this area relies on the availability of data on service *wait times* for a centralized organization, such as airline flight data (Balakrishna et al. 2010, Ravizza et al. 2014, Simaiakis and Balakrishnan 2015) or emergency rooms in hospitals (Ibrahim and Whitt 2011a, Ang et al. 2015). Although this approach addresses an important service outcome that can impact consumer behavior (Armony and Maglaras 2004, Allon et al. 2011a, Jouini et al. 2011, Xu et al. 2016, Dong et al. 2019), it only measures one possible dimension of service quality.

A related research stream has focused on collecting feedback directly from customers measured along multiple dimensions using surveys (Grönroos 1994, Soteriou and Chase 2000, Bitran et al. 2008). For example, the SERVQUAL instrument (Parasuraman et al. 1988) proposes 22 questionnaire items to measure the gap between expectations and perceptions at the time of service. The results from SERVQUAL can then be correlated with economic outcomes to understand the consequences of service quality. Despite the adoption of these surveys (Metters and Marucheck 2007), their applicability to different service industries has been questioned (Cronin and Taylor 1992, Roth and Jackson 1995, Roth and Menor 2003). Creating reliable surveys requires adapting to varying service types where customers weight service attributes differently (Kettinger and Lee 2005, Ladhari 2009). Furthermore, surveys are time-consuming (Cronin and Taylor 1992); provide limited validity across similar providers, making benchmarking difficult (Roth and Jackson 1995); do not capture individual heterogeneity (Chase and Apte 2007); and do not scale (Roth and Menor 2003). These issues can be addressed by alternative methods of capturing service quality at scale that are nevertheless nonintrusive and economical (Voss et al. 2008). We argue that social media data provides an opportunity to measure service quality.

2.2. Opportunities in Social Media and Online Reviews to Examine Quality Indicators

Online reviews have emerged as a viable alternative source of information on consumer experience for potential customers and service providers (Forman et al. 2008, Mudambi and Schuff 2010). Since the advent of platforms like Yelp and Amazon, the ability of online reviews to influence consumer behavior

has become well-established (Dellarocas 2003). Recent consumer surveys reinforce this notion; in 2017, "97% of U.S. consumers read online reviews for local businesses," and "68% of consumers said that positive reviews made them more likely to use a business" (BrightLocal 2017).

Online reviews are particularly influential in the services context. Garnering informative online reviews and interacting with consumers online are critical behaviors for service providers to engage customers (van Doorn et al. 2010). Further, positive online reviews are correlated with firm performance indicators, such as sales and consumer choice (Chen and Xie 2008), loyalty (Yoo et al. 2013), trust (Awad and Ragowsky 2008), and brand image (Dellarocas 2003). These studies explain economic or behavioral outcomes by using one or a combination of numerical measures, such as dispersion, valence, and volume (Dellarocas and Narayan 2006), across multiple contexts like television ratings (Godes and Mayzlin 2004), movie box office sales (Liu 2006, Duan et al. 2008), and healthcare (Gao et al. 2012). Within the restaurant sector, in particular, Anderson and Magruder (2012) show that star ratings on Yelp have a direct effect on the availability of reservations for restaurants. In an influential paper, Luca (2016) provides evidence for the impact of review star ratings on restaurant revenues.

In most of these studies, researchers use the numerical data associated with online reviews. However, a number of potential issues arise from using these variables to represent the consumer experience. By compressing complex customer engagements to a single number, product or service quality is assumed to be one-dimensional, which is unrealistic (Archak et al. 2011). Further, individual preferences are highly heterogeneous, and a single number may not convey the same information to everyone. Finally, numerical star ratings have been shown to have a significant bias (Li and Hitt 2008, Chen and Lurie 2013). For these reasons, a more granular approach to capture service quality that also utilizes text data should be preferred (van Doorn et al. 2010). Recent work has focused attention on review text by considering word counts, readability indices, and sentiment (Decker and Trusov 2010, Archak et al. 2011, Netzer et al. 2012). However, a more in-depth analysis of review text to measure service quality effectively is still warranted.

Natural language processing offers rigorous and systematic tools for extracting valuable information embedded in the textual content of reviews. Specifically, it offers techniques for topic modeling and sentiment analysis where the general aims are to identify the discussion themes that are present in the set of documents and the tone surrounding these themes. In this paper, we use NMF as our primary methodology to process review text. We choose this method due to its interpretability, simplicity, and similarity with principal component analysis (PCA), a commonly used technique. NMF allows the identification of semantic service-related themes in the text of online reviews. To demonstrate the robustness of our overall findings, we also use the ABSA technique provided by Brody and Elhadad (2010) that identifies simultaneous estimation of service themes as well as sentiment across these themes in each review, described in online Appendix B. We establish the relationship between the semantic service quality themes and survival next.

2.3. Survival and the Strategic Importance of Quality

The literature is full of correlational studies linking service quality to operational and strategic benefits for the firm (Soteriou and Chase 2000, Chase and Apte 2007). It has been asserted that: "in the long run, the most important single factor affecting a business unit's performance is the quality of its products and services relative to those of competitors" (Buzzell and Gale 1987, p. 7). Consistent with this viewpoint, we argue that the semantic themes of service quality should be associated with measures of service provider performance.

The relationship between service quality and performance has been hypothesized in several theoretical frameworks in the services literature. Rust et al. (1995) treat service quality as an investment on the part of the provider, defining "return on quality" to describe the mechanisms by which high quality yields positive financial returns. The service-profit chain model (Sasser et al. 1997, Heskett et al. 2008) postulates that service quality and customer satisfaction lead to customer loyalty, which in turn leads to improved financial performance, and has received empirical validation (Sasser et al. 1997, Soteriou and Chase 2000, Roth and Menor 2003). A third approach comes from Roth and Jackson (1995), who use the capabilities, service quality, and performance triad (C-SQ-P) to argue that providers deliver better quality because they are able to employ their capabilities more effectively, compared with competitors.

Building on this research, we focus on one important outcome, survival. Survival remains understudied in the services literature despite its relevance and importance (Kalnins and Mayer 2004, Rosenzweig et al. 2011). In the restaurant sector, specifically, survival has been examined through case studies (Parsa et al. 2005, 2011) and industry-level analyses using the U.S. Bureau of Labor Statistics data (Luo and Stark 2015). However, research linking provider-level service quality to survival is absent. Clearly, several factors within the service providers' ecological environment, such as competitive intensity, institutional support, and regulation, can affect survival

(Castrogiovanni 1991, Carroll and Khessina 2005). However, specific attributes about the service provider, such as superior service quality, should enable the provider to thrive (Gu et al. 2011, Rosenzweig et al. 2011). Service quality represents an important intermediary variable between firm capabilities (Roth and Menor 2003) and performance. We test this broad hypothesis using the semantic dimensions extracted from online reviews as a final form of validation.

3. Methodology

3.1. Research Context

For this study, we focus on restaurants in Washington, DC. Restaurants represent a common service context where online reviews are used extensively, and survival is a critical outcome (Gu et al. 2011, Lu et al. 2013). As a first step, we identify all open restaurants in the area, as of December 2013, using the Washington, DC, municipal city database. This provides us with a master list of more than 2,000 restaurants that are "going concerns" during the time of data collection. We then assemble a comprehensive list of restaurant closures in the DC area, drawing from two data sources. First, we collect search results of restaurants reported as closed on Yelp and Foursquare. Second, we collect reports of restaurant closings from Eater and Gayot, two sites dedicated to local culinary events. In total, this data collection effort results in a list of 575 restaurants that closed between 2005 and 2013. Out of these 575 restaurants, some were near DC but located in Virginia and Maryland, which were removed, leaving us with 446 restaurants located inside DC. Finally, we manually confirmed that these 446 restaurants truly closed and did not simply change locations. While this set of closed restaurants may not be comprehensive, it represents a wide cross-section of cuisines, segments, and locations within DC. Further, the data allows us to approximate the closing period of the restaurant to the nearest quarter.

Next, we gather online reviews for this set of restaurants. Although much work in this area relies on web-scraping, it may be seen as a violation of terms and conditions for some platforms. Instead, we use the application programming interface (API) provided by Yelp, the leading platform for restaurant reviews. Yelp provides reviews on restaurants dating back to 2004. To use their API, we first submitted an application to the firm in 2013 when the project was first conceived, with a detailed description of our planned analyses and research questions. We ensured that our work was consistent with the API's terms of use, in particular, the stipulation that users were allowed to create a database of Yelp content, create new metrics, and perform analysis of Yelp content, as long as our work did not imply or lead to any commercial use of the data. Once our application was approved, we were granted a private key to download the data.

Furthermore, the terms of use also state the manner in which the data should be downloaded. Specifically, Yelp mandated that users cannot impact the stability of Yelp's servers and are limited to 5,000 daily server calls. We were able to stay within this limit while downloading the relevant review data for our sample of restaurants in a timely manner.[1]

Additionally, we omit reviews that are tagged by Yelp as being suspicious or fraudulent following recent work on Yelp data (Byers et al. 2012, Luca and Zervas 2016). This process resulted in over 35,000 reviews for the closed restaurants and a total of 130,000 reviews for all restaurants. Each review is time-stamped and associated with a star rating, allowing us to create a panel data set of reviews over time. In addition, we also collect information associated with each restaurant, such as the location, price point, cuisine, and parking. These features represent control variables in our models and are described in Table 1. Next, we describe the process by which the semantic themes pertaining to service quality were identified.

3.2. Extracting Service Themes from Text

With review text, we follow standard preprocessing procedures in text analysis: transforming text to lowercase and removing words composed of less than three characters, common words called stop words, and stemming words. The general objective of preprocessing is to focus on meaningful words by removing uninformative ones and to keep the number of unique terms that appear in the corpus from becoming extremely large, one of the main computational challenges in text mining. Examples of stop words include "the," "and," and "of." Stemming refers to the process of removing suffixes, so words like values, valued, and valuing are all replaced with "valu." We use the Porter stemming algorithm, which iteratively applies linguistic rules to identify and remove suffixes. Porter stemming is a standard algorithm implemented in most text mining software, including the "tm" (text mining) package (Feinerer and Hornik 2012) within R (R Core Team 2000), used for our analysis.

After preprocessing the text, as our primary methodology for extracting semantic dimensions, we utilize NMF, a well-established topic modeling technique for extracting such dimensions (often called topics, factors, or themes in related fields) from text data. The NMF process represents each review with a set of numerical covariates that capture service themes within the review text. It resembles PCA for numerical data, as both techniques are based on a matrix factorization with the same algebraic form as singular value decomposition (SVD). NMF differs from SVD in the constraints that are placed on the matrix decomposition.

Table 1. Variable Descriptions

Variable	Description
$Closure_{it}$	1 if restaurant i is closed in time period t; 0 otherwise
$meanrating_{it}$	Average rating of reviews for restaurant i in time period t
$QualityOverall_{it}$	NMF measure for the overall experience in restaurant i in time period t
$WaitTimes_{it}$	NMF measure for the reliability and *wait times* in restaurant i in time period t
$FoodQuality_{it}$	NMF measure for the *food quality* in restaurant i in time period t
$Responsiveness_{it}$	NMF measure for the service *responsiveness* in restaurant i in time period t
$Atmosphere_{it}$	NMF measure for the atmosphere in restaurant i in time period t
$Pricepoint_i$	Price point for restaurant i
WL_{it}	Average word count of reviews for restaurant i in time period t
$readability_{it}$	Average Simple Measure of Gobbledygook (SMOG) readability index of reviews for restaurant i in time period t
$comp\_meanrating_{it}$	Average mean rating for restaurant i's competitors in time period t
$comp\_numreviews_{it}$	Number of reviews for restaurant i's competitors in time period t
$numcompetitors_{it}$	Number of competitors for restaurant i in time period t
$Cuisine_i$	Set of binary variables indicating whether each of 16 cuisines is listed in the cuisine type for restaurant i (restaurants can have multiple cuisines)
Loc_i	Set of binary variables indicating the zip code of restaurant i
$OtherChars_i$	Set of binary variables describing 15 other characteristics for restaurant i, such as payment method, parking, attire, group-friendly, kid-friendly, waiter, Wi-Fi, and alcohol

Instead of ortho-normality constraints of the factors that are defined by SVD, NMF requires nonnegativity, that is, every element of the matrix factortion must be greater than or equal to zero. Such nonnegativity constraints have been shown to produce more meaningful results for text data specifically (Xu et al. 2003, Shahnaz et al. 2006, O'Callaghan et al. 2015), in addition to several other forms of data (Lee and Seung 1999, Berry et al. 2007). Consequently, when combined with linear regression models, which is our goal, the NMF procedure yields the best set of results when compared with alternative topic modeling techniques. In online Appendices A and B, we provide a detailed discussion of text-based NMF in addition to a comparison of several popular topic modeling methods.

Similar to PCA, it is necessary to evaluate how many semantic themes may emerge from applying NMF to the data set of reviews first. The process of identifying the most suitable number of themes, referred to as cross-validation, suggests five service themes underlying the review text (described in online Appendix A). Once the themes are identified, each review is represented with numerical scores for each of the extracted themes (akin to PCA component loadings), and each theme has a list of rank-ordered keywords (shown in Table 2) that we inspect to understand the semantic meaning behind the extracted themes. In many such text-based settings, it is up to the researcher to examine the underlying text associated with a specific theme and evaluate their relevance. Thus, manual examination of the extracted themes and the application of domain knowledge in this process is considered best practice within machine learning and text analysis literature (Sebastiani 2002, Arora et al. 2012). In our case, as shown in Table 3, phrases from top reviews that load heavily on each service theme are interpretable and consistent with an underlying service dimension. Note that all reviews have semantic theme loadings on all five themes (similar to PCA); the reviews shown in Table 3 are selected because they load heavily on the relevant theme, thereby making it easier for the data scientist to interpret and name the underlying themes appropriately. Based on the phrases and reviews

Table 2. Top Words Associated with Each Semantic Component

Quality overall	Wait times	Responsiveness	Food quality	Atmosphere
food	place	order	good	place
good	bar	server	veri	crowd
place	order	came	great	happi
like	time	ask	dish	bar
order	one	dish	nice	beer
friend	drink	menu	flavor	neighborhood
time	wait	waiter	chicken	like
great	minut	minut	restaur	food
nice	hour	tabl	sauc	drink
service	ask	meal	food	music

Table 3. Reviews Associated with the Five Semantic Components

Reviews that load most heavily onto quality overall
"Food was okay. My shrimp tempura roll was good, but the donburi wasn't. The tempura ice cream was my favorite part of the meal. The service was pretty good. Our server was a genuine sweetheart so I might go again for the rolls and the service. Pretty place too."
"The atmosphere of the place was kind of weird. The food wasn't all that impressive for Thai food in the DC area. . .If you get 'Beef Red Curry,' you kind of expect more than 4 small pieces of beef. . .The service though was outstanding. The server was always around for water/drink refills and was very nice."
"We weren't seated in the main dining area (that's for the highrollers). . . Palena was a good meal overall, not stressful as some good restaurants tend to get around the busiest dinner hours."
Reviews that load most heavily onto wait times
"It took them 30 minutes to make 2 burgers."
"It's just as good as toki underground, and better yet there is not a ridiculous two hour wait. . .you order at the counter and they prepare it right away. you get your food within 5 min. . . it kept me coming back again and again."
"We went on a weekday night, and we were told there was a 15 minute wait. The host scooped us out of the bar not even a minute later with an available table."
"We got there earlier than our reservation time (7PM on a Thursday) and were able to be seated right away."
Reviews that load most heavily onto responsiveness
"my medium well burger came out medium rare, and they put bacon on my wife's veggie burger (she did not ask for this obviously)."
"I really, really liked the vending machine that was dispensing beer and cigarettes – you had me at hello. I also thought the bartender was super friendly and accommodating'
"Not only did the server have great difficulty comprehending the neatly written break-down, he came back and told us that we were $1 short of the "suggested gratuity" shown on the receipt."
"Service was prompt and pleasant"
"My fiancé and I were gracious to be sat quickly at a one of the last tables in the full dining room set for four. . .. Notably, different members of the staff delivered plates and shared a little bit about what we were about to enjoy (without aimlessly listing off ingredients)"
"Our waitress Rosalin(?) was nice enough, but very very confident and almost seemed like she was acting out a scene as she told us a bit about the restaurant and theme."
Reviews that load most heavily onto food quality
"The shrimp were perfect, really perfect, I had to resist stealing more from my friend. I enjoyed the fried oysters in cornmeal too. The shrimps were just really flavorful, just really good."
"The delicate rings of squid – so exquisitely supple save for one or two pieces – were served on a velvety polenta, fire-roasted tomato fondue, and fresh pesto. A perfectly portioned and divine way to begin what was about to be our oceanic adventure. . .The lobster was excruciatingly tender and sweet, paired with a clarified lemon-herb butter. The oysters were enormous and juicy, paired with a mouth-puckering mignonette. The jumbo shrimp were bigger than jumbo, delicate meat executed perfectly, paired with a spicy cocktail sauce. The mussels were wonderfully plump and meaty."
"We consider it one of the best meals we've ever had so far. . .. Course after course titillated and awed, I loved every minute of it."
"everyone was really pleased with their food. . .. The tempura was quite good. The soup was a little bit too salty and the dumpling really greasy, but overall, everything was good."
"Corned beef sliders - good. Heavy, yummy. . .. Shepherd's pie . . .was just ground lamb and potatoes, . . .Lemon something for dessert - looked very yummy. Eh. Even the shortbread was eh."
Reviews that load most heavily onto atmosphere
"The atmosphere of the place was kind of weird. There are rocks on the ceiling and it was a little bit too dark. I guess they were going for a cave theme, but the important question is WHY???"
"It smells greasy. Atmosphere is lacking, but since this is more a take out place for lunch time, it doesn't have to have atmosphere. Then again, why not work a bit at it. Easiest first step - turn on some music, not much, not loud, just a little basic background. I understand New Orleans is famous for its music. . .and the guys working here might also like that."
"what i like about this place is you get great food for a very reasonable price in a laid back, low key, and personable environment."

associated with these phrases, we name the five themes: *overall quality*, *wait times*, *food quality*, *responsiveness*, and *atmosphere*.

3.2.1. Comparing NMF with Other Methodologies. Our overall goal is to show how text-based analysis can help provide a useful and scalable measure of service quality. Our main results in the following sections are presented using NMF. However, there are alternative methodologies available for this analysis. As a test of consistency across text mining implementations, we also consider a second form of text-based analysis called ABSA, which aims to distinguish explicitly between the amount of discussion on an estimated topic (i.e., aspect) and the corresponding sentiment for each topic in each review. Sentiment remains an important component of any text processing and provides useful information to readers. In NMF, sentiment is estimated separately. However, in ABSA, the procedure provides both the identification of the topic and the associated sentiment. Moreover, ABSA methods typically analyze the text at the sentence level, which has been empirically shown to be advantageous in certain settings (Brody and Elhadad 2010).

For example, treating individual sentences as documents can result in recovered topics that correspond more strongly to service quality (e.g., wait times), whereas treating entire reviews as documents (as NMF does) can recover topics that distinguish more between reviews (e.g., cuisines). On the other hand, the joint estimation of topic-sentiment structures and finer resolution of ABSA create higher computational and implementation costs relative to NMF. Indeed, there are standard libraries for NMF, whereas most ABSA methodologies require the data scientist to develop specialized routines.

The literature in ABSA methodology has several techniques developed to address specific contexts and data. For example, Pontiki et al. (2015, 2016) and Wang and Liu (2015) created supervised methods that assume that one has access to training data in the form of reviews that are annotated at the sentence level with aspects and sentiment information. Without such training data, the methodology described involves a high up-front cost to label the sentences. In Yelp, reviews are not labeled or annotated. Thus, to match our specific setting where we observe the overall star rating and the text of the review, we implement two popular ABSA techniques: (i) Wang et al. (2010), which utilizes both the text and overall rating of the review for estimation; (ii) Brody and Elhadad (2010), which performs the estimation using only the text of the reviews. Performing these analyses allows us to compare results from NMF and ABSA. We find that NMF and the two ABSA techniques produce similar service quality themes and have similar goodness-of-fit indicators. In terms of external validation explaining Zagat ratings and restaurant closure (described in detail in the next section), we find that NMF has slightly superior performance compared with ABSA. Overall, the body of results confirms that our findings are robust to the choice of the text mining algorithm while showing support for the overall strategy of using text to measure elements of service quality.

3.3. External Validation of Service Quality Themes in a Controlled Laboratory Setting

We now turn to the validation of the semantic dimension scores associated with each restaurant. These validation tests are necessary to establish that the extracted dimensions do capture recognizable dimensions of service quality, provided by existing service quality standards in the industry, for a given restaurant at a given time. Furthermore, we would like to determine if the information captured by the semantic scores are comparable to the process by which actual consumers may respond to online reviews. Therefore, as part of this strategy, we conduct two sets of validation tests, described in this and the following sections.

The first validation test pertains to evaluating consumers' responses to experimental data that is drawn from the NMF-based semantic themes, effectively testing whether consumers recognize differences in reviews that score high (or low) on the semantic dimensions, absent any other source of variation. Based on this validation, we can link the review text and the underlying semantic themes behaviorally to how the average consumer perceives and responds to the reviews provided for a restaurant. Therefore, we test if individual reviews that load highly on the *overall quality* dimension, that is, containing a highly detailed description of the restaurant's *overall quality*, are perceived as different from other reviews by consumers in controlled settings. If a strong case for the sensitivity (i.e., customers correctly classify reviews that load highly in this specific dimension) and specificity (i.e., customers correctly classify reviews that do not load highly in this specific dimension) can be made, we can conclude that the semantic themes are indeed capturing relevant and discernible information about service quality. We conducted a series of laboratory experiments on Amazon Mechanical Turk (MTurk) to examine whether customers accurately classify reviews that load highly or lowly on the semantic dimensions for a fictional restaurant. The use of MTurk as a platform for running large-scale experiments has gained wide acceptance in the social sciences. MTurk's participant pool is considered reliable for experimental research (Goodman et al. 2013), represents the broader population (Buhrmester et al. 2011), and can generate high-quality results (Ipeirotis 2010). In settings like ours, where primary data are generally hard to procure except in a university laboratory setting under very artificial conditions, MTurk's experimental platform is particularly useful.

Before running the final experiment on MTurk, we conducted a pilot test using a small set of undergraduate subjects from a large North American university. We ran these experiments progressively for each of the five dimensions identified using the NMF procedure described above. The pilot was meant to ensure that the stimuli used in the eventual experiments were valid and reliable and is highly encouraged in experimental research (De Vaus 2001, Van Teijlingen et al. 2001). For the first three dimensions (*overall quality*, *wait times*, and *responsiveness*), the pilots were successful, indicating that pilot subjects were able to classify reviews loading highly (lowly) on *overall quality* on these dimensions correctly. We also received feedback from the respondents in terms of improving the stimuli once the pilot was completed. After making the suggested changes in experimental design, we proceeded to recruit subjects for the eventual large-scale MTurk experiment for the three semantic dimensions. For the two remaining

dimensions (*atmosphere* and *food quality*), the pilots did not provide evidence of valid treatments. The subjects were unable to clearly detect differences in reviews loading highly (lowly) on these dimensions. While online Appendix C provides more details on the pilot results, these two dimensions do represent the fourth and fifth topics extracted. As is the case with PCA, subsequent components extracted explain lesser variance in the source data and therefore provide fewer diagnostic signals compared with the earlier identified components. It is possible that this issue makes it difficult for pilot subjects to classify reviews pertaining to these dimensions clearly. This is not to infer that the *atmosphere* and *food quality* are not relevant dimensions of service quality; rather, the NMF procedure is not able to generate these dimensions cleanly from review text. We, therefore, do not conduct large-scale experiments for these semantic dimensions. The inability to validate these two dimensions experimentally at scale remains a limitation.

In each experiment, subjects were shown a specific online review and asked about their purchase intention, that is, how likely they were to visit the restaurant, given the reviews. The objective here is to gauge whether reviews that loaded highly on the focal semantic dimension influenced decision making. Purchase intention is defined as the probability that customers intend to purchase a particular product or service (Chang and Wildt 1994, Grewal et al. 1998) and is a reliable predictor of revenues in the marketing literature (Mittal and Kamakura 2001, Gupta and Zeithaml 2006), ultimately leading to higher odds of survival for service firms (Kandampully and Suhartanto 2000). We note that these experiments are not designed to fully replicate the decision-making process by which customers choose restaurants; indeed, doing so is impossible within a laboratory experiment. Instead, the experiments are designed to test whether subjects can recognize elements of service quality within the text of online reviews and respond as expected.

For each of the three dimensions tested (*overall quality*, *wait times*, *responsiveness*), 200 subjects were assigned to one of four 2 (dimension: high vs. low) × 2 (numerical rating: high vs. low) between-subjects' conditions. The stimuli were developed by randomly selecting an actual review from our data set that loaded highly/lowly on each of the semantic dimensions. Participants were asked to read the selected review about the restaurant and then assess the quality of the fictitious restaurant as well as their likelihood to visit the restaurant. An important experimental design parameter here is the number of reviews to show the subject. Although it is theoretically possible to show subjects a set of reviews instead of one review, we encounter issues associated with the order of reviews, spillover effects from one review to the next, and the need to retain consistency across reviews pertaining to the same fictional restaurant. From an experimental viewpoint, showing multiple reviews raises many issues of complexity that are not directly relevant here. Therefore, in the interest of simplicity and parsimony, we show each MTurk subject only one specific review, ensuring that each review within the experiment was approximately the same length and other information on the restaurant provided was identical (except for review text). We were careful to use the Yelp interface, using identical font and color schemes. Details of the procedures, data, and analysis of the experiments are provided in online Appendix C.

We use a binomial test to gauge whether the classifications of the displayed online reviews provided by MTurk subjects were statistically different from chance. The binomial test compares the observed distribution of data to the null hypothesis that the MTurk subjects respond purely by chance, that is, 50% probability of providing a high/low classification. The results, which are provided in full in online Appendix C, show that when subjects are exposed to the "high" case of all three semantic dimensions (*overall quality*, *wait times*, and *responsiveness*), they provide the right classificatory response (93%, 90%, and 85%, respectively). However, for the "low" treatments, the subject responses are less accurate. The responses to *overall quality* and *responsiveness* are statistically distinguishable from chance, but those to *wait times* are not. As a first observation, we note that subjects are able to distinguish between the reviews that load highly on specific semantic dimensions and those that have low loadings. However, there is an interesting asymmetry in how well subjects are able to classify high and low loadings—reviews that score highly on semantic dimensions are more cleanly classified.

One possible explanation for this effect comes from the dynamics observed in online reviews in general (Dellarocas 2003, Anderson and Magruder 2012), which tend to display an upward bias in ratings. Thus, for a 1–5 scale, the realized mean rating is often above 4. It is thus possible that clearly positive semantic content in the reviews is easier to observe and classify by humans since this is consistent with what is expected. On the other hand, when the text is not as positive or presents ambiguity to the consumer, the source of the ambiguity and the resulting quality perception is not easily established. Thus, our subjects are able to classify reviews that load highly on the semantic dimensions. We also find that such reviews are significantly associated with the subjects' purchase intention. When aggregated up to a more expansive set of reviews on a restaurant being viewed

by many thousand potential customers, the cumulative effects on demand for the restaurant are discernible here. These effects are weaker in the case of reviews that have low loadings on the semantic dimensions. Although the observed asymmetry does not nullify the potential relationship between the semantic dimensions and restaurant closure, it represents an interesting topic for future research. We now move to the second external validation test: comparing the semantic dimensions to alternative measures of service quality.

3.4. External Validation of Service Quality Themes with Zagat Ratings

Do the extracted semantic themes of service quality match up with externally provided and accepted measures of service quality? To answer this question, we need to compare our data to an alternative source. We use restaurant ratings from Zagat, a national restaurant review firm founded in 1979 to measure the quality of restaurants in New York City. Since then, the firm has expanded domestically and internationally and now operates in 38 U.S. cities (Zagat 2010). Zagat provides one of the most influential scores measuring restaurant quality, with the Wall Street Journal describing it as "the gastronomic bible" (Efrati 2011). Zagat's rating procedure involves a 30-point survey aggregated across three dimensions: food, decor, and service. Volunteer correspondents, who mostly remain anonymous, register with Zagat and rate and submit surveys of restaurants over a given year. Meals are not reimbursed by Zagat, so correspondents are not financially tied to the merchant or the publisher. Zagat editors, who are active local food critics (Zagat 2010), then aggregate the information from correspondents and assign scores along each of the three dimensions. Editors also select quotes from customers to include in the review.

Although Zagat utilizes three dimensions of the dining experience (food, decor, and service) compared with our more granular five (*overall quality*, *wait times*, *food quality*, *responsiveness*, and *atmosphere*), they appear to measure largely similar themes. For example, Zagat's service aligns with our *wait times* and *responsiveness*. Similarly, Zagat's food aligns reasonably with our *food quality*, while Zagat's decor is similar to our *atmosphere*. Our semantic themes, generated via decentralized feedback from average consumers, appear to be consistent with an industry standard.

In order to validate this quantitatively, we compare the semantic dimension scores to the scores published by Zagat for the same restaurant in the same year. Substantial correlation between the two sets of scores would indicate that our measures of service quality are indeed in agreement with an industry standard. Since Zagat ratings for the period of our study are only available in book form, we first coded the information for the Washington, DC, Zagat books for the years 2005–2013 except for 2009, since no book was published that year. As with the Yelp data set, we selected only restaurants within DC. Restaurants across the Zagat and Yelp data sets were matched using the restaurant name, address, and telephone number. Approximately 97% of the restaurants in Zagat were successfully matched to Yelp. With this composite data set, we used seemingly unrelated regression (SUR) (Zellner 1962, Greene 2012) to assess agreement with Zagat measures. Specifically, we estimate the system of equations shown below, where i indexes restaurants and t indexes years:

$$\begin{aligned} Zagat_{Food_{it}} &= b_{0i} + \beta_1 QualityOverall_{it} + \beta_2 \\ &\quad \times FoodQuality_{it} + \beta_3 t + \epsilon_{1it}, \end{aligned} \tag{1}$$

$$\begin{aligned} Zagat_{Decor_{it}} &= b_{0i} + \beta_1 QualityOverall_{it} + \beta_2 \\ &\quad \times FoodQuality_{it} + \beta_3 Atmosphere_{it} \\ &\quad + \beta_4 t + \epsilon_{2it}, \end{aligned} \tag{2}$$

$$\begin{aligned} Zagat_{Service_{it}} &= b_{0i} + \beta_1 QualityOverall_{it} + \beta_2 \\ &\quad \times FoodQuality_{it} + \beta_3 Atmosphere_{it} \\ &\quad + \beta_4 WaitTime_{it} + \beta_5 Responsiveness \\ &\quad + \beta_6 t + \epsilon_{3it}. \end{aligned} \tag{3}$$

We include fixed effects to account for restaurant heterogeneity and time trends. We fit the model using data from 899 restaurants that appear on both Yelp and Zagat during at least one year. The estimated SUR model results are shown in Table 4 and have appropriate levels of goodness of fit, with a McElroy R^2 for the system of 0.65. Interestingly, *overall quality* is significantly correlated with all three Zagat dimensions. *Wait times* and *responsiveness* are both correlated with Zagat's service dimension (the negative coefficient for *wait times* is expected since higher wait times are associated with worse service). Finally, *food quality* and *atmosphere* are correlated with Zagat's food and decor dimensions, respectively. These significant correlations show that our semantic scores are capturing similar elements as Zagat's dimensions. In online Appendix B, we replicate this exercise using the scores from ABSA and find similar results, with a somewhat weaker model fit.

In summary, after identifying five themes reflecting service quality dimensions from online review text, our validation tests show that the themes may be used to evaluate restaurant service quality reliably. Through the use of MTurk experiments as well as Zagat ratings, we find that the dimensions (a) provide a reasonable classification of the text of online reviews that is externally validated by humans subjects in MTurk in a controlled environment and (b) are strongly correlated with an external source of service quality based on a defined

Table 4. SUR Results for Zagat Ratings

Variable	Equation (1) Zagat's food	Equation (2) Zagat's decor	Equation (3) Zagat's service
QualityOverall	2.65 (0.67)***	2.05 (0.93)*	2.64 (1.01)**
FoodQuality	3.05 (1.01)**	5.36 (16.01)	3.55 (10.28)
Atmosphere		2.11 (0.53)***	0.50 (1.90)
WaitTimes			−2.99 (0.38)**
Responsiveness			1.87 (0.61)**
Restaurant fixed effects	Included	Included	Included
n (observations)	7,192	7,192	7,192
Groups (restaurants)	899	899	899
Multiple R^2	0.62	0.59	0.67

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

process that provides information, albeit only on a yearly basis. We now discuss our third set of validation tests: are the semantic dimensions able to predict firm-level outcomes better than numerical ratings alone? We test for the relationship, if any, between the semantic dimensions and restaurant survival, arguably the most critical form of validation since this addresses an important and relevant economic outcome. Although many factors affect restaurant survival (Parsa et al. 2011), we examine the marginal effects of service quality, measured using the semantic dimensions and aggregated across online reviews for a restaurant, on the odds of a restaurant's survival in the short run.

4. Econometric Models and Estimation

Our data set includes information about restaurants that closed as well as the reviews that the restaurant received on Yelp in the periods prior to closing. Our goal is to estimate a regression model where the dependent variable is the restaurant's status at a point in time (open or closed) and the independent variables are the review-based scores from the received reviews prior to that time period. Therefore, we construct a panel data set where the unit of observation is the restaurant–time period. Within each unit of observation, we include restaurant characteristics, the reviews the restaurant received in that time period as well as semantic scores emerging from these received reviews and the status of the restaurant. For the baseline analysis, we use quarterly time periods, focusing on reviews published in that quarter. Several time-invariant restaurant characteristics remain constant. We also exclude restaurant–time periods occurring before the first restaurant's review, after restaurant closure, and with no new reviews, resulting in an unbalanced panel data set. We average the semantic scores on each dimension across all new reviews in each restaurant–time period. In online Appendix E, we also estimate models where the time period is six months, with variables appropriately redefined, and find consistent results.

We specify the general form of the regression equation using semantic themes and restaurant status. The variable Closure_{it} equals 1 if restaurant i has closed at time t and 0 otherwise. We include a set of control variables that capture heterogeneity across restaurants and over time. These variables include fixed restaurant characteristics Chars_i, such as the price point of restaurant i, type of cuisine, and availability of alcohol. We also include time-varying review variables Chars_{it}, such as the number of reviews received during time period t for restaurant i and the average star rating received by restaurant i in time period t. Summary statistics for these variables are shown in Table 5. A correlation matrix is available upon request. Although we could include aggregate sentiment (averaged over reviews), we find that it is highly correlated with the average star rating, as observed previously by Cao et al. (2011); we therefore only include rating in our main models. In online Appendix E, alternative models, including sentiment, are presented and show similar results. In online Appendix B, we consider alternative models using ABSA scores, which incorporate sentiment into semantic dimensions; these models result in similar coefficient estimates but weaker model fit. Thus, we include only the aggregate semantic scores from NMF, NMF_{it}.

Table 5. Summary Statistics

Variable	Mean	Standard deviation
Meanrating	3.316	0.921
Numreviews	5.173	5.085
OverallQuality	6.610	6.218
WaitTimes	5.709	7.837
FoodQuality	6.021	7.557
Responsiveness	4.938	7.235
Atmosphere	6.007	6.295
Pricepoint	1.774	0.638
WL	120.627	68.394
Readability	8.875	2.085
Comp_meanrating	3.313	0.363
Comp_numreviews	2.898	2.061
Num_competitors	39.203	25.246

We also control for review length and readability scores of each review, as suggested by Ghose and Ipeirotis (2011). Since we are modeling survival, we include variables from prior research within the restaurant industry that may drive survival. Location is one such factor. Washington, DC, is divided into 12 zip codes, so we include fixed effects for each location (Loc_i). Furthermore, restaurants exist in highly competitive environments. Although competitive intensity may be defined along different dimensions, we capture it in a simple and parsimonious manner. For each restaurant i, we define competition as other restaurants within the same price point and location and include the average rating and number of restaurants in this set. Although this captures one form of competitive intensity, it provides no information on how much traffic these competitors actually generate. Therefore, we use the number of reviews received in each time period by these competitors and use the average as the measure of competition as well. Intuitively, this captures the extent to which the restaurant operates in an environment with high or low traffic to comparable restaurants. These are collectively denoted Comp_{it}. For the sake of interpretability, all independent variables are scaled and centered. We also lag the independent variables by one time period, thereby reducing the chances of reverse causality (discussed in the next section). The baseline model for estimation is the logistic regression defined by

$$\log\left(\frac{P(\text{Closed}_{i,t+1})}{1-P(\text{Closed}_{i,t+1})}\right) = b_{i0} + \beta_0 + \gamma\text{Chars}_i + \delta\text{Chars}_{it} + \alpha\text{NMF}_{it} + \rho\text{Loc}_i + \zeta\text{Comp}_{it} + \epsilon_{i,t+1}, \quad (4)$$

where γ and δ are vectors containing the sets of coefficients corresponding to restaurant and review characteristics, respectively, α contains the coefficients corresponding to the NMF variables, and ρ and ζ are vectors containing the set coefficients corresponding to location and competition. We use a logistic regression specification wherein the dependent variable is the log odds of closure in the next time period. In the next section, we start by describing the construction of our data set using matching techniques to minimize omitted variable bias. Then, we describe the estimation of the models and end the section by addressing issues of identification, reverse causality, and endogeneity that may introduce bias into coefficient estimates along with a summary of other robustness checks.

4.1. Coarsened Exact Matching

First, we address the potential of omitted variable bias, whereby factors that are unobservable to the researcher may influence both service quality and the closure decision. Although omitted variable bias is difficult to rule out without perfect randomization, matching methods are often used in observational studies to reduce bias in the estimated treatment effects due to unobserved variables (Dehejia and Wahba 2002). Specifically, by matching restaurants that close to others that do not but are otherwise similar, one can make causal inferences about the effect of the semantic themes on closure (Stuart 2010). We used coarsened exact matching (CEM) as it has several beneficial statistical properties compared with other methods such as propensity-score-matching (PSM) and distance-based matching (Iacus et al. 2009). Next, we describe how we create a matched sample using CEM. As a robustness check, we also perform matching using PSM and naive matching. We find consistent results using all three matching techniques.

For each closed restaurant, the objective of CEM is to identify a set of restaurants that remained open but match closely in terms of other covariates. The actual process of matching consists of three main steps: coarsening each matching variable, sorting observations into strata defined by the coarsened variables, and discarding unbalanced strata. These steps are described in detail in Iacus et al. (2012). We employ the progressive coarsening method outlined in Iacus et al. (2009) to select the matching variables, and we use mean rating, number of reviews, price point, cuisine, and descriptive characteristics (excluding review text) of the restaurant provided on Yelp, consistent with prior work (Lu et al. 2013). Since we have a longitudinal data set, we match each closed restaurant with open restaurants in the time period exactly before the focal restaurant closed.

The CEM procedure results in all 446 closed restaurants being matched to 605 open restaurants (out of 2,021 candidates). We utilize $1:k$ matching, which has been shown to be desirable (Stuart 2010). To evaluate the balance in the case and control samples, we examined an imbalance table as recommended by Iacus et al. (2009) and found the two samples to be balanced sufficiently. A comparison of the matched and treated samples (Table 6) shows similar values on all numerical covariates, indicating a good match.

4.2. Longitudinal Generalized Linear Models

We can now use this composite sample of control and treated restaurants to estimate the theoretical model

Table 6. Comparing Case and Control Samples

Sample	*n*	*Meanrating*	*Numreviews*	*Pricepoint*
Case	446	3.26	54.80	1.76
Control	605	3.31	59.50	1.65

postulated earlier. We use a generalized linear mixed model specification (GLMER; McCulloch and Neuhaus 2006) with a binomial family using a logit link and a random intercept (equivalent to a panel logit model). We estimate different variations on the baseline model, with and without the semantic themes, starting with a regression specification that assumes that, in each period, a restaurant chooses whether to exit the market. Our focus is on evaluating the incremental explanatory power the semantic themes provide, representing service quality, given other covariates, as follows.

Model 1 only includes restaurant and review characteristics:

$$\log\left(\frac{P(\text{Closed}_{i,t+1})}{1-P(\text{Closed}_{i,t+1})}\right) = b_{i0} + \beta_0 + \gamma\text{Chars}_i + \delta\text{Chars}_{it} + \rho\text{Loc}_i + \zeta\text{Comp}_{it} + \epsilon_{i,t+1}. \quad (5)$$

Model 2 includes restaurant and review characteristics as well as the semantic variables:

$$\log\left(\frac{P(\text{Closed}_{i,t+1})}{1-P(\text{Closed}_{i,t+1})}\right) = b_{i0} + \beta_0 + \gamma\text{Chars}_i + \delta\text{Chars}_{it} + \alpha\text{NMF}_{it} + \rho\text{Loc}_i + \zeta\text{Comp}_{it} + \epsilon_{i,t+1}. \quad (6)$$

Model 3 is the same as Model 2, except that the sample of restaurants i are restricted to the set of closed restaurants Ω. Model 3 allows us to examine the robustness of the results to the matching process. In addition to Models 1–3, we also fit two models using the GLMER but with added fixed effects for each restaurant and year. The added fixed effects provide more conservative coefficient estimates while accounting for restaurant-specific heterogeneity. Model 4 applies an exchangeable correlation structure, whereas Model 5 applies an autoregressive correlation structure with 1 lag (AR(1); McCulloch and Neuhaus 2006). The correlation structure refers to how observations from the same restaurant, that is, within a group, are assumed to be correlated and affect standard errors but not coefficient estimates. The exchangeable correlation structure assumes that every pair of observations in a group has the same correlation, whereas the AR(1) structure assumes that observations closer in time strongly correlate to each other, as observed in prior work in online reviews (Moe and Trusov 2011). As a robustness test, we estimated models with lagged variables using more than one-period lags for the semantic components. The current-period variables remained significant, whereas deeper lags were not significant. These results are available in online Appendix E and are important in assessing model reliability (Liang and Zeger 1986). The specification for Model 4, including restaurant and year fixed effects, is

$$\log\left(\frac{P(\text{Closed}_{i,t+1})}{1-P(\text{Closed}_{i,t+1})}\right) = b_{i0} + \beta_0 + \theta_{Year_t} + \gamma\text{Chars}_i + \delta\text{Chars}_{it} + \alpha\text{NMF}_{it} + \epsilon^{exc}_{i,t+1}, \quad (7)$$

where Year_t is the year portion of the year-quarter time period t and ranges from 2004 to 2013. Similarly, Model 5 has the following specification, including restaurant and year fixed effects:

$$\log\left(\frac{P(\text{Closed}_{i,t+1})}{1-P(\text{Closed}_{i,t+1})}\right) = b_{i0} + \beta_0 + \theta_{Year_t} + \gamma\text{Chars}_i + \delta\text{Chars}_{it} + \alpha\text{NMF}_{it} + \epsilon^{AR(1)}_{i,t+1}. \quad (8)$$

The results from the estimation of each of these models are shown in Table 7. We first see that a number of control variables are significant in influencing the odds of restaurant closure. Restaurants featuring certain cuisines (e.g., American food) have a higher probability of closure. Further, as expected, the average star rating (highly rated restaurants) and the number of reviews (higher traffic) are negatively associated with the probability of closure ($p < 0.001$). These results are reflective of prior work describing the importance of online reviews in the service context (Dellarocas 2003).

Regarding the semantic dimensions (service quality) from review text, we observe a range of p-values, coefficient magnitudes, and signs across the five variables. The coefficient of the first semantic dimension, *overall quality*, is highly significant ($p < 0.001$ in all four models) and strongly correlated with closure. Furthermore, as expected based on our discussion earlier, the coefficient for the *wait times* dimension is positively associated with closure ($p < 0.01$ in all four models). The dimension titled *responsiveness* is marginally significant ($p < 0.1$) and is negatively associated with closure. Strikingly, the marginal effects of *overall quality* and *wait times* are larger than that of the average star rating in all models (since all variables are scaled and centered), indicating that these semantic dimensions might be more influential than the average star rating. Interestingly, not all NMF-based themes are significant: *food quality* and *atmosphere* are not significant in any of our models.

Importantly, the addition of the semantic dimensions leads to a significant improvement in model fit,

Table 7. GLMER Coefficient Estimates

	DV = Closure$_{i,t+1}$				
Variable	Equation (5) Model 1 base model	Equation (6) Model 2 semantic variables	Equation (6) Model 3 closed sample	Equation (7) Model 4 fixed effects (EX)	Equation (8) Model 5 fixed effects (AR-1)
Intercept	–5.29 (0.23)***	–5.83 (0.25)***	–3.72 (0.19)***	–4.05 (0.16)***	–4.11 (0.17)***
Meanrating	–0.12 (0.05)*	–0.15 (0.04)***	–0.11 (0.04)*	–0.14 (0.06)*	–0.14 (0.06)*
Numreviews	–2.07 (0.18)***	–2.12 (0.19)***	–2.02 (0.20)***	–1.89 (0.22)***	–1.91 (0.18)***
QualityOverall		–0.65 (0.17)***	–0.54 (0.18)**	–0.79 (0.10)***	–0.77 (0.14)***
WaitTimes		0.17 (0.06)**	0.16 (0.06)*	0.23 (0.08)**	0.28 (0.09)***
FoodQuality		–0.06 (0.06)	–0.07 (0.06)	–0.07 (0.08)	–0.07 (0.13)
Responsiveness		–0.17 (0.08)*	–0.13 (0.07) _	–0.08 (0.07)	–0.02 (0.15)
Atmosphere		0.02 (0.06)	0.01 (0.05)	0.11 (0.19)	0.17 (1.01)
Pricepoint	–0.05 (0.09)	–0.03 (0.12)	–0.05 (0.06)		
Wordlength	–0.34 (0.07)***	0.17 (0.17)	0.45 (0.17)		
Readability	–0.02 (0.05)	–0.03 (0.04)	0.01 (0.04)		
comp_meanrating	–0.05 (0.06)	–0.07 (0.11)	–0.17 (0.11)		
comp_numreviews	0.48 (0.06)***	0.79 (0.28)**	1.37 (0.01)***		
num_competitors	–0.02 (0.07)	–0.01 (0.01)	–0.01 (0.01)		
Cuisine FE	Included	Included	Included	Included	Included
Zip code FE	Included	Included	Included	Included	Included
OtherChars FE	Included	Included	Included	Included	Included
Restaurant-year FE	Omitted	Omitted	Omitted	Included	Included
n (observations)	16,515	16,515	4,787	16,515	16,515
Groups (restaurants)	1,035	1,035	437	1,035	1,035
AIC	3,822.9	3,677.4	2,558.0		
BIC	3,933.2	3,812.6	2,701.2		
QIC				2,478.3	2,393.1

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

as shown by the differences in the Bayesian information criterion (BIC) and Akaike information criterion (AIC) values. Based on the guidelines provided by Burnham and Anderson (2004) for model selection, the differences in AIC scores between Model 2 and Model 1 in Table 7 are significant and provide substantial support for the proposed model that augments numerical variables with the semantic scores from NMF.

The results from *overall quality* and *responsiveness* obtain negative coefficients, showing that restaurants with reviews that clearly indicate engagement on *overall quality* and *responsiveness* are less likely to close (i.e., Closure$_{it}$ = 0). Furthermore, the positive coefficient for *wait times* is consistent with validation tests conducted with Zagat, which showed that, in many cases, long *wait times* tend to be negatively associated with customer perceptions. These tests are also consistent with the results from the MTurk experiments discussed earlier, in that reviews that load heavily on these dimensions elicit a higher propensity to actually visit the restaurant, all else being equal. In general, we see a consistent pattern of results, showing that the semantic dimensions are reflective of service quality that can be measured across time and used for comparisons across restaurants.

The results from Model 3, which only includes closed restaurants, are similar to those from Model 2, based on the matched sample. This similarity first shows that the matching process has not radically modified any inferences that may be drawn from the analysis and, second, provides a degree of robustness to the baseline results. Finally, we note that including the semantic dimensions in the analysis significantly improves the overall fit of the model—a 16% and a 10% reduction in the AIC and BIC of Model 2 compared with Model 1, respectively. The AIC and BIC reductions imply noticeably better model fit and show the value of the NMF-based semantic information captured from the review text, even in the presence of traditional numerical variables (Burnham and Anderson 2004).

4.3. Discussing Endogeneity, Reverse Causality, and Other Robustness Checks

In this section, we address some potential issues with the econometric models we estimate above. The main independent variables in our analysis are the five semantic dimensions based on NMF. A critical assumption in our regression model here is that service quality, gleaned from the restaurant's reputation on Yelp, is associated with the restaurant's survival. However, it is possible that the restaurant's reputation on Yelp is what drives the restaurant's survival instead. From an identification perspective, it is not possible for us to separate these two effects effectively with secondary data, which remains a limitation in this context. With respect to the five semantic themes we use in our analysis, it is also possible to argue that these themes are endogenous, that is, they

are conditioned on forward-looking expectations of outcomes (e.g., closure), thereby introducing bias into the estimation of their marginal effects on survival. However, unlike survey-based methodologies that rely on a small subset of customers (Parasuraman et al. 1988), several hundred reviewers provide text in a decentralized and independent manner. Therefore, it is unlikely that a large subset of reviewers have similar forward-looking expectations of restaurant closure. This is one of the potential strengths of crowd-sourced content: individual, systematic biases may be significantly lower than those found in contexts with smaller, more targeted sets of respondents (Dellarocas 2003). Thus, we expect endogeneity from forward-looking expectations of survival to be less of a concern here.

An associated concern in our models is reverse causality between service quality and closure, that is, the decision to close a restaurant may be followed by a reduction in service quality before the restaurant actually closes, which is, in turn, reflected in online reviews. The presence of reverse causality would require that, following a decision to close, restaurateurs remain in operation during the quarter following the closure decision. To ascertain the likelihood of this, we interviewed several members of the National Restaurant Association and were informed that most restaurateurs prefer to "down shutters" as rapidly as possible once the actual closure decision has been made. This suggests that reverse causality is unlikely to play a major role in our analysis. As a robustness check, we estimate our models using six-month time periods rather than quarters and find consistent results (online Appendix E). Additionally, we investigate the evolution of the semantic scores for restaurants that close in order to shed light on the extent to which reverse causality may be at play. Specifically, we examine if service quality does indeed decline suddenly beginning several months prior to restaurant closure. We find that our semantic scores are relatively flat 1.5 months prior to closure. Thus, we conclude that the potential for biased estimates due to reverse causality is unlikely in our model.

Beyond specific issues around endogeneity and reverse causality, we conduct an expansive set of robustness tests on the models described in the previous section. Though they are mentioned elsewhere in the paper, we list them out here for completeness. First, we have thus far viewed restaurant closure as a binary event. Treating closure as an outcome of a duration model (Lin and Wei 1989), we estimate a Cox proportional survival model. Second, deeper time lags of the semantic dimensions, to capture longer-term effects of service quality, were estimated. Third, we control for restaurant age in the analysis (Parsa et al. 2005). Fourth, we estimated models controlling for review sentiment. The ABSA methodology reported in the Appendix online Appendix B also addresses this issue in some detail, albeit with a weaker model fit. Fifth, we use alternative matching methods (PSM and naive matching) to create the matched sample and re-estimate the models. Although individual coefficients change, the results remain qualitatively similar across these models. Finally, in some contexts, the researcher may be interested in predicting survival. We consider the accuracy of a prediction model using receiver operating characteristic (ROC) curves and provide this analysis in the Appendix online Appendix D. The prediction model shows which service quality scores may be used to predict which restaurants are likely to close in the near future, providing a clear benefit for managers and investors alike.

5. Discussion and Managerial Implications

Service quality has remained a critical antecedent of performance in the service industry for many years (Chase and Apte 2007). However, considerable research on service quality notwithstanding, two significant gaps still exist. First, the measurement of service quality in a systematic and scalable manner that can be extended beyond a particular type of service provider remains a limitation in the literature (Metters and Marucheck 2007). Traditional approaches using surveys and operational metrics, such as measuring customer wait times, have significant limitations in terms of scale and face validity, particularly at a time when customers are increasingly focused on experiences (Pine and Gilmore 1999, Buell et al. 2016) and engagement (Voss et al. 2008). This sentiment is succinctly captured by Soteriou and Zenios (1999, p. 1225), "In order to answer the 'how' questions we need to address 'what' questions first. What are the operational characteristics of a service that translate to customers' high levels of quality?" Second, although researchers have linked service quality to operational outcomes at the level of a specific process or customer interaction (Soteriou and Chase 2000, Heskett et al. 2008, Kostami and Ward 2008), the macro-level implications for the service provider in terms of survival have not been studied in detail. There are many reasons why a service provider may fail; in the context of competitive services, providing relatively poorer service quality should be a significant factor driving failure.

In our work here, we address these two gaps. First, we tackle the issue of service quality measurement by using online reviews for restaurants from Yelp as the primary source of data. Using text analytic methods, such as NMF and ABSA, we extract a set of service quality dimensions from the text of the reviews. Our approach to extracting insight from text is not only scalable but also allows for comparisons across

service providers. Further, our approach benefits from the large corpora of review text since the ability to identify dimensions of interest improves with corpus size (Blei 2012). Therefore, the strategy of using text as a measure of service quality promises to get better over time as review corpora grow in size. To validate these service quality dimensions, we employ a variety of validation tests: we first ensure that reviews that are associated with the extracted dimensions match the perceptions of human subjects who read these reviews and subsequently evaluate restaurants in a series of controlled experiments. We then determine the extent to which these semantic dimensions correlate with service quality industry standards (e.g., Zagat).

After validating the dimensions, we then estimate models that link service quality, measured by semantic dimensions, to the survival of restaurants in Washington, DC. We show that the semantic dimensions are correlated with restaurant exit across the years 2005–2013. Of the five service quality dimensions extracted, two are consistently significant predictors of restaurant closure, while a third is marginally significant. These dimensions continue to be significant even after controlling for review star rating, showing that there is useful information about the service experience available in the text. A recent report by the National Restaurant Association (2016) stated, "Simply put: online reviews can help or break your business" through its influence on the consumer's decision to choose a restaurant; we see significant support for this perspective.

We note certain limitations in the research reported here. First, we do not provide a comprehensive model of restaurant closure; we do not have access to other factors that may be influential, such as sales data. Second, although we use several forms of topic modeling, including NMF and ABSA (Brody and Elhadad 2010), other contexts may benefit from alternative topic modeling methods. Our results suggest that NMF is particularly advantageous since it is computationally more efficient, is more intuitive since it is comparable to PCA, and provides a strong econometric fit to predict closure. As newer text analysis algorithms emerge in the future, a deeper examination of the specific trade-offs between these algorithms is warranted.

Our work also provides several insights to managers related to service industries. First, although online reviews inexpensively provide access to data on customer experiences (Cao et al. 2011), manually processing this data is infeasible. Using the strategy that we propose, restaurant managers can quickly identify dominant themes to understand their performance and compare against competitors better. Further, platforms such as Yelp can use our approach to provide merchants and customers multidimensional scores of quality using review text that might more accurately describe the service quality of a merchant.

Acknowledgments

The authors thank the department editor, associate editor, and review team for their constructive suggestions. Special thanks to Ishwarya Sethuraman for her help in data collection and to Kislaya Prasad and Joseph P. Bailey for their advice.

Endnote

[1] See the full terms of use of the Yelp API at https://www.yelp.com/developers/api_terms.

References

Allon G, Bassamboo A, Gurvich I (2011a) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

Allon G, Federgruen A, Pierson M (2011b) How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing Service Oper. Management* 13(4):489–507.

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

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2015) Accurate emergency department wait time prediction. *Manufacturing Service Oper. Management* 18(1):141–156.

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

Armony M, Maglaras C (2004) Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52(4):527–545.

Armony M, Shimkin N, Whitt W (2008) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.

Arora S, Ge R, Moitra A (2012) Learning topic models – going beyond SVD. *2012 IEEE 53rd Annual Sympos. Foundations Comput. Sci.*, 1–10.

Awad NF, Ragowsky A (2008) Establishing trust in electronic commerce through online word of mouth: An examination across genders. *J. Management Inform. Systems* 24(4):101–121.

Balakrishna P, Ganesan R, Sherry L (2010) Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. *Transportation Res. Part C: Emerging Tech.* 18(6):950–962.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* 61(1):39–59.

Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Statist. Data Anal.* 52(1):155–173.

Bitran GR, Ferrer JC, Rocha e Oliveira P (2008) OM forum—managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing Service Oper. Management* 10(1):61–83.

Blei DM (2012) Probabilistic topic models. *Commun. ACM* 55(4): 77–84.

BrightLocal (2017) Local consumer review survey 2017 and the impact of online reviews. Accessed December 1, 2017, https://www.brightlocal.com/learn/local-consumer-review-survey/.

Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. Kaplan RM, ed. *Human Language Technologies:*

The 2010 Annual Conf. Amer. Chapter Assoc. Comput. Linguistics (Association for Computational Linguistics, Stroudsburg, PA), 804–812.

Buell RW, Campbell D, Frei FX (2016) How do customers respond to increased service quality competition? *Manufacturing Service Oper. Management* 18(4):585–607.

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psych. Sci.* 6(1):3–5.

Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33(2):261–304.

Buzzell RD, Gale BT (1987) *The PIMS Principles: Linking Strategy to Performance* (Simon and Schuster, New York).

Byers JW, Mitzenmacher M, Zervas G (2012) Daily deals: Prediction, social diffusion, and reputational ramifications. *Proc. 5th ACM Internat. Conf. Web Search Data Mining*, 543–552.

Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50(2):511–521.

Carroll GR, Khessina OM (2005) The ecology of entrepreneurship. Alvarez SA, Agarwal R, Sorenson O, eds. *Handbook of Entrepreneurship Research* (Springer, Boston), 167–200.

Castrogiovanni GJ (1991) Environmental munihcence; A theoretical assessment. *Acad. Management Rev.* 16(3):542–565.

Chang TZ, Wildt AR (1994) Price, product information, and purchase intention: An empirical study. *J. Acad. Marketing Sci.* 22(1):16–27.

Chase RB, Apte UM (2007) A history of research in service operations: What's the big idea? *J. Oper. Management* 25(2): 375–386.

Chen Y, Xie J (2008) Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Sci.* 54(3):477–491.

Chen Z, Lurie NH (2013) Temporal contiguity and negativity bias in the impact of online word of mouth. *J. Marketing Res.* 50(4): 463–476.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Cronin JJ, Taylor SA (1992) Measuring service quality: A reexamination and extension. *J. Marketing* 56(3):55–68.

Cui R, Li J, Zhang D (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci.* 66(3):1071–1094.

Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. *Rev. Econom. Statist.* 84(1): 151–161.

Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Sci.* 49(10):1407–1424.

Dellarocas C, Narayan R (2006) A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statist. Sci.* 21(2):277–285.

De Vaus DA (2001) *Research Design in Social Research* (Sage, Thousand Oaks, CA).

Dong J, Yom-Tov E, Yom-Tov GB (2019) The impact of delay announcements on hospital network coordination and waiting times. *Management Sci.* 65(5):1969–1994.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter? — an empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Efrati A (2011) Zagat deal extends Google's influence. *Wall Street Journal* (September 9), https://www.wsj.com/articles/SB10001424053111904836104576558600549181370.

Feinerer I, Hornik K (2012) tm: Text mining package. R package version 0.5-7.1. Accessed July 29, 2020, http://tm.r-forge.r-project.org/.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3): 291–313.

Gao GG, McCullough JS, Agarwal R, Jha AK (2012) A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. *J. Medical Internet Res.* 14(1):e38.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Goldstein SM, Johnston R, Duffy J, Rao J (2002) The service concept: The missing link in service design research? *J. Oper. Management* 20(2):121–134.

Goodman JK, Cryder CE, Cheema A (2013) Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decision Making* 26(3):213–224.

Greene W (2012) *Econometric Analysis* (Pearson Education, Upper Saddle River, NJ).

Grewal D, Krishnan R, Baker J, Borin N (1998) The effect of store name, brand name and price discounts on consumers' evaluations and purchase intentions. *J. Retailing* 74(3):331–352.

Grönroos C (1994) From scientific management to service management: A management perspective for the age of service competition. *Internat. J. Service Indust. Management* 5(1):5–20.

Gu B, Park J, Konana P (2011) Research note—the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Inform. Systems Res.* 23(1):182–196.

Gupta S, Zeithaml V (2006) Customer metrics and their impact on financial performance. *Marketing Sci.* 25(6):718–739.

Hall J, Porteus E (2000) Customer service competition in capacitated systems. *Manufacturing Service Oper. Management* 2(2):144–165.

Heineke J, Davis MM (2007) The emergence of service operations management as an academic discipline. *J. Oper. Management* 25(2):364–374.

Heskett J, Jones T, Loveman G, Sasser E, Schlesinger L (2008) Putting the service-profit chain to work. *Harvard Bus. Rev.* (July-August):118–129.

Iacus SM, King G, Porro G (2009) CEM: software for coarsened exact matching. *J. Statist. Software* 30(13):1–27.

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Anal.* 20(1):1–24.

Ibrahim R, L'Ecuyer P (2012) Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing Service Oper. Management* 15(1):72–85.

Ibrahim R, Whitt W (2011a) Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Prod. Oper. Management* 20(5):654–667.

Ibrahim R, Whitt W (2011b) Wait-time predictors for customer service systems with time-varying demand and capacity. *Oper. Res.* 59(5):1106–1118.

Ipeirotis PG (2010) Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads* 17(2):16–21.

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* 13(4):534–548.

Kalnins A, Mayer KJ (2004) Franchising, ownership, and experience: A study of pizza restaurant survival. *Management Sci.* 50(12): 1716–1728.

Kandampully J, Suhartanto D (2000) Customer loyalty in the hotel industry: The role of customer satisfaction and image. *Internat. J. Contemporary Hospitality Management* 12(6):346–351.

Kettinger WJ, Lee CC (2005) Zones of tolerance: Alternative scales for measuring information systems service quality. *Management Inform. Systems Quart.* 29(4):607–623.

Kim SH, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing Service Oper. Management* 16(3):464–480.

Kostami V, Ward AR (2008) Managing service systems with an offline waiting option and customer abandonment. *Manufacturing Service Oper. Management* 11(4):644–656.

Ladhari R (2009) A review of twenty years of SERVQUAL research. *Internat. J. Quality Service Sci.* 1(2):172–198.

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Inform. Systems Res.* 19(4):456–474.

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22.

Lin DY, Wei LJ (1989) The robust inference for the cox proportional hazards model. *J. Amer. Statist. Assoc.* 84(408):1074–1078.

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

Lu X, Ba S, Huang L, Feng Y (2013) Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Inform. Systems Res.* 24(3):596–612.

Luca M (2016) Reviews, reputation, and revenue: The case of yelp.com. Preprint, submitted March 15, https://ssrn.com/abstract=1928601.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Sci.* 62(12): 3412–3427.

Luo T, Stark PB (2015) Nine out of 10 restaurants fail? Check, please. *Significance* 12(2):25–29.

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

McCulloch CE, Neuhaus JM (2006) *Generalized Linear Mixed Models* (John Wiley & Sons, Hoboken, NJ).

Metters R, Marucheck A (2007) Service management—academic issues and scholarly reflections from operations management researchers. *Decision Sci.* 38(2):195–214.

Mittal V, Kamakura WA (2001) Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *J. Marketing Res.* 38(1):131–142.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.

Mudambi S, Schuff D (2010) What makes a helpful online review? A study of customer reviews on amazon.com. *Management Inform. Systems Quart.* 34(1):185–200.

National Restaurant Association (2016) Restaurant industry forecast. Accessed December 1, 2017, http://www.restaurant.org/Home/old_files/Forecast-2016.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

O'Callaghan D, Greene D, Carthy J, Cunningham P (2015) An analysis of the coherence of descriptors in topic modeling. *Expert Systems Appl.* 42(13):5645–5657.

Parasuraman A, Zeithaml VA, Berry LL (1988) Servqual: A multiple-item scale for measuring consumer perc. *J. Retailing* 64(1):12–40.

Parsa HG, Self J, Sydnor-Busso S, Yoon HJ (2011) Why restaurants fail? Part II - The impact of affiliation, location, and size on restaurant failures: Results from a survival analysis. *J. Foodservice Bus. Res.* 14(4):360–379.

Parsa HG, Self JT, Njite D, King T (2005) Why restaurants fail. *Cornell Hotel Restaurant Admin. Quart.* 46(3):304–322.

Pine BJ, Gilmore JH (1999) *The Experience Economy: Work Is Theatre & Every Business a Stage* (Harvard Business Press, Brighton, MA).

Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I (2015) Semeval-2015 task 12: Aspect based sentiment analysis. *Proc. 9th Internat.Workshop Semantic Evaluation (SemEval 2015)*, 486–495.

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Mohammad AS, Al-Ayyoub M, et al (2016) Semeval-2016 task 5: Aspect based sentiment analysis. *Proc. 10th Internat. Workshop Semantic Evaluation (SemEval-2016)*, 19–30.

R Core Team (2000) R language definition. R Foundation for Statistical Computing, http://www.R-project.org/.

Ramdas K, Williams J, Lipson M (2013) Can financial markets inform operational improvement efforts? Evidence from the airline industry. *Manufacturing Service Oper. Management* 15(3): 405–422.

Ravizza S, Chen J, Atkin JAD, Stewart P, Burke EK (2014) Aircraft taxi time prediction: Comparisons and insights. *Appl. Soft Comput.* 14(C):397–406.

Rosenzweig ED, Laseter TM, Roth AV (2011) Through the service operations strategy looking glass: Influence of industrial sector, ownership, and service offerings on B2B e-marketplace failures. *J. Oper. Management* 29(1):33–48.

Roth AV, Jackson WE (1995) Strategic determinants of service quality and performance: Evidence from the banking industry. *Management Sci.* 41(11):1720–1733.

Roth AV, Menor LJ (2003) Insights into service operations management: A research agenda. *Prod. Oper. Management* 12(2):145–164.

Rust RT, Zahorik AJ, Keiningham TL (1995) Return on quality (ROQ): Making service quality financially accountable. *J. Marketing* 59(2):58–70.

Sasser WE, Schlesinger LA, Heskett JL (1997) *Service Profit Chain* (Simon and Schuster, New York, NY).

Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput. Surv.* 34(1):1–47.

Shahnaz F, Berry MW, Pauca VP, Plemmons RJ (2006) Document clustering using nonnegative matrix factorization. *Inform. Process. Management* 42(2):373–386.

Simaiakis I, Balakrishnan H (2015) A queuing model of the airport departure process. *Transportation Sci.* 50(1):94–109.

Soteriou A, Chase RB (2000) A robust optimization approach for improving service quality. *Manufacturing Service Oper. Management* 2(3):264–286.

Soteriou A, Zenios SA (1999) Operations, quality, and profitability in the provision of banking services. *Management Sci.* 45(9): 1221–1238.

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. *Statist. Sci.* 25(1):1–21.

van Doorn J, Lemon KN, Mittal V, Nass S, Pick D, Pirner P, Verhoef PC (2010) Customer engagement behavior: Theoretical foundations and research directions. *J. Service Res.* 13(3):253–266.

Van Teijlingen E, Hundley V (2001) The importance of pilot studies. *Soc. Res. Update* 35(4):1–4.

Veeraraghavan S, Debo L (2008) Joining longer queues: Information externalities in queue choice. *Manufacturing Service Oper. Management* 11(4):543–562.

Voss C, Roth AV, Chase RB (2008) Experience, service operations strategy, and services as destinations: Foundations and exploratory investigation. *Prod. Oper. Management* 17(3):247–266.

Wang B, Liu M (2015) Deep learning for aspect-based sentiment analysis. Report, Stanford University, Palo Alto, CA.

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. Rao B, Krishnapuram B, eds. *Proc. 16th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 783–792.

Wyckoff DD (2001) A Cornell quarterly classic article: New tools for achieving service quality. *Cornell Hotel Restaurant Admin. Quart.* 42(4):25–38.

Xu K, Chan CW (2016) Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing Service Oper. Management* 18(3):314–331.

Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. *Proc. 26th Annual Internat. ACM SIGIR Conf. Research Development Inform. Retrieval*, 267–273.

Xu Y, Armony M, Ghose A (2016) The interplay between online reviews and physician demand: An empirical investigation. Preprint, submitted August 8, https://ssrn.com/abstract=2778664.

Ye Q, Law R, Gu B (2009) The impact of online user reviews on hotel room sales. *Internat. J. Hospitality Management* 28(1): 180–182.

Yelp (2017) About us. Accessed December 1, 2017, https://www.yelp.com/about.

Yoo CW, Sanders GL, Moon J (2013) Exploring the effect of e-WOM participation on e-loyalty in e-commerce. *Decision Support Systems* 55(3):669–678.

Zagat (2010) *Zagat 2010: Washington, D.C.–Baltimore Restaurant Survey Update* (Zagat Survey, New York).

Zeithaml VA, Berry LL, Parasuraman A (1996) The behavioral consequences of service quality. *J. Marketing* 60(2):31–46.

Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* 57(298):348–368.