International Journal of Hospitality Management 44 (2015) 120–130

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhosman

What can big data and text analytics tell us about hotel guest experience and satisfaction?

Zheng Xiang[a,*], Zvi Schwartz[b], John H. Gerdes Jr.[c], Muzaffer Uysal[a]

[a] Department of Hospitality and Tourism Management, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA
[b] Department of Hotel, Restaurant & Institutional Management, University of Delaware, Newark, DE 19716, USA
[c] Department of Integrated Information Technology, College of Hospitality, Retail, & Sport Management, University of South Carolina, Columbia, SC, USA

ARTICLE INFO

Keywords:
Big data
Text analytics
Guest experience
Satisfaction
Hotel management

ABSTRACT

The tremendous growth of social media and consumer-generated content on the Internet has inspired the development of the so-called big data analytics to understand and solve real-life problems. However, while a handful of studies have employed new data sources to tackle important research problems in hospitality, there has not been a systematic application of big data analytic techniques in these studies. This study aims to explore and demonstrate the utility of big data analytics to better understand important hospitality issues, namely the relationship between hotel guest experience and satisfaction. Specifically, this study applies a text analytical approach to a large quantity of consumer reviews extracted from Expedia.com to deconstruct hotel guest experience and examine its association with satisfaction ratings. The findings reveal several dimensions of guest experience that carried varying weights and, more importantly, have novel, meaningful semantic compositions. The association between guest experience and satisfaction appears strong, suggesting that these two domains of consumer behavior are inherently connected. This study reveals that big data analytics can generate new insights into variables that have been extensively studied in existing hospitality literature. In addition, implications for theory and practice as well as directions for future research are discussed.

1. Introduction

Social media and consumer-generated content on the Internet continue to grow and impact the hospitality industry (Browning et al., 2013; Xiang and Gretzel, 2010). The tremendous growth of these data-generating sources has inspired the development of new approaches to understanding social/economic phenomena in a variety of disciplines (Wood et al., 2013; George et al., 2014). The so-called big data analytics approach emphasizes and leverages the capacity to collect and analyze data with an unprecedented breadth, depth, and scale to solve real-life problems (Ginsberg et al., 2009; Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013). In the hospitality field there is a growing interest in utilizing user-generated data to gain insights into research problems that have not been well understood by conventional methods (e.g., Ye et al., 2009a; Yang et al., 2013). Indeed, big data analytics opens the door to numerous opportunities to develop new knowledge to reshape our understanding of the field and to support decision making in the hospitality industry. However, while a handful of studies have employed new data sources to tackle important research problems, they were conducted on an *ad hoc* basis and the application of the big data analytics approach in hospitality is yet to be well developed and established.

The goal of this study is to explore and demonstrate the utility of big data analytics by using it to study core hospitality management variables that have been extensively studied in past decades. Specifically, hotel guest experience and satisfaction have long been a topic of interest because it is widely recognized that they contribute to customer loyalty, repeat purchases, favorable word-of-mouth, and ultimately higher profitability (see Oh and Parks, 1997). Particularly, the hotel industry is highly competitive in that hotel firms offer essentially homogeneous products and services, which drive the desire of hotels to distinguish themselves among their competitors. As such, guest satisfaction has become one of the key measures of a hotel's effectiveness in outperforming others. Since the 1970s a plethora of studies has been conducted with the aim to understand the components and antecedents of guest satisfaction (e.g., Choi and Chu, 2001; Hunt, 1975; Mattila and O'Neill, 2003; Oh, 1999; Pizam et al., 1982; Su, 2004; Wu and Liang, 2009).

* Corresponding author. Tel.: +1 5402313262.
E-mail addresses: philxz@vt.edu (Z. Xiang), zvi@UDel.edu (Z. Schwartz), gerdes@mailbox.sc.edu (J.H. Gerdes Jr.), samil@vt.edu (M. Uysal).

http://dx.doi.org/10.1016/j.ijhm.2014.10.013
0278-4319/© 2014 Elsevier Ltd. All rights reserved.

While this line of research offers a variety of perspectives on guest satisfaction, the vast majority of existing studies primarily relied upon conventional research techniques such as consumer surveys or focus group interviews to gauge what leads to guest satisfaction. As such, whether we can develop novel and meaningful insights into these building blocks of hospitality management using big data analytics becomes an intriguing research question.

This study employed one of the most important types of consumer-generated content, i.e., online customer reviews of hotel properties, to understand hotel guest experience and its relationships with guest satisfaction. Text analytics was applied to first deconstruct a large quantity of customer reviews collected from Expedia.com and then examine its association with hotel satisfaction ratings. Thus, the analytics approach aimed to gain insights into the nature and structure of guest experience expressed when a customer gave a specific satisfaction rating for the hotel he/she has stayed in. This paper is organized as follows: following the introduction, the subsequent section reviews literature on the big data analytics approach and hotel guest experience and satisfaction. Research questions are formulated with the focus on using online customer reviews to enrich our understanding of these constructs. The methodology section details data collection and the text analytical approach utilized to answer the research questions. Findings are then presented and discussed. Finally, the study's contributions to literature and practice as well as directions for future research are discussed.

2. Research background

2.1. Big data analytics and business intelligence

Big data is being generated through many sources including Internet traffic (e.g., clickstreams), mobile transactions, user-generated content, and social media as well as purposefully captured content through sensor networks, business transactions, and many other operational domains such as bioinformatics, healthcare, and finance (George et al., 2014). Big data analytics aims to generate new insights that can meaningfully and, oftentimes in real time, complement traditional statistics, surveys, and archival data sources that remain largely static. The classic example of big data analytics is the pioneer study using Google search queries to detect epidemic diseases in the society (Ginsberg et al., 2009). As demonstrated by the study, big data analytics leads to a profound epistemological change that reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality (Boyd and Crawford, 2012). As such, big data analytics can be seen as a new research paradigm, rather than a uniform method, that may utilize a diverse set of analytical tools to make inferences about reality using large data. Importantly, although big data analytics does not preclude hypothesis testing, it is often applied to explore novel patterns or predict future trends from the data (Aiden and Michel, 2014). While it is widely accepted as a new approach to knowledge creation, there has been recently voice of concerns about the potential pitfall of spurious correlations and thus calls for theory-based approaches to big data analytics (Boyd and Crawford, 2012; Marcus and Davis, 2014)

One of the application areas of growing importance is the so-called business intelligence in that big data analytics can be used to understand customers, competitors, market characteristics, products, business environment, impact of technologies, and strategic stakeholders such as alliance and suppliers. Many examples and cases have been cited to illustrate the applications of big data analytics to discover and solve business problems (Mayer-Schönberger and Cukier, 2013). Mining social media and consumer-generated content has attracted much attention for their value as public and community data (George et al., 2014). For instance, research has demonstrated that online consumer reviews can be used to predict product quality (Finch, 1999), stock market volatility (Antweiler and Frank, 2004; Schumaker and Chen, 2009) and box office sales in the movie industry (Duan et al., 2008). It has been found that online news postings have sufficient linguistic content to be predictive of a firm's earnings and stock returns (Tetlock et al., 2008). More recently, Ghose and Ipeirotis (2011) used text content and reviewer characteristics to estimate the helpfulness and economic impact of online hotel product reviews. Abrahams et al. (2012) devised a technique to detect automobile defects through online consumer discussion forums. Moreover, it has been shown that marketing tools such as product recommender systems can be developed based upon the mining of consumer-generated content in combination with other data sources (Ghose et al., 2012).

Due to the volume and unstructured nature of social media and consumer generated content, opinion mining and sentiment analysis, i.e., the so-called text analytics, plays an important role in big data analytics. Indeed, opinion mining and sentiment analysis is considered well-suited to various types of market intelligence applications (Pang and Lee, 2008). Sentiment-analysis technologies for extracting opinions from unstructured human-authored documents can be excellent tools for handling many business intelligence tasks including reputation management, public relations, tracking public viewpoints, as well as market trend prediction. Broadly speaking, sentiment analysis and opinion mining denote the same techniques that are derived from and based upon natural language processing (NLP), information retrieval (IR), information extraction (IE), and artificial intelligence (AI). Typical tasks of sentiment analysis include (1) finding documents relevant for a specific topic or purpose; (2) pre-processing collected documents, e.g., tokenizing documents into single words and extracting relevant information from them; and (3) identifying the sentiment surrounding the product or company (Schmunk et al., 2013). In comparison with the broader scope of text mining approach, sentiment analysis may be considered a special type of text mining with the focus on identification of subjective statements and contained opinions and sentiments, particularly in consumer-generated content on the Internet.

2.2. Hotel guest experience and satisfaction

Hotel guest satisfaction is a complex human experience within a hospitality service setting. The study of guest satisfaction was initiated as early as the 1970s. Different definitions of guest satisfaction have emerged. Hunt (1975) considers satisfaction as an evaluation on which the customers have experienced with the services is at least as good as it is supposed to be, while others (e.g., Oliver, 1981) define customer satisfaction as an emotional response to the use of a product or service. Oh and Parks (1997) postulate that satisfaction involves cognitive and affective processes, as well as other psychological and physiological influences. A commonly used definition of customer satisfaction adopts a disconfirmation perspective of consumer satisfaction/dissatisfaction, suggesting that satisfaction is the result of the interaction between a consumer's pre-purchase expectation and post-purchase evaluation (Engel et al., 1990). In the tourism literature, various perspectives have been employed to conceptualize the concept of tourist satisfaction including the expectation/disconfirmation paradigm, the equity view, the norm view, as well as the perceived overall performance (see Yoon and Uysal, 2005 for a comprehensive review).

From the managerial point of view, it is, perhaps, more important to understand the components or antecedents of hotel guest satisfaction. For example, it has been conceptualized that the hotel product consists of several levels. That is, the core product, i.e., the hotel room, deals exactly with what the customer receives from the purchase. Besides, the hotel product also includes facilitating, supporting, and augmenting elements which concern with, for example, how the customer receives from the purchase, the interactions with service providers and other customers, as well as necessary conditions (e.g., the front desk) which provide access to the core product and numerous value-added products and services (Kotler et al., 2006). The hotel product can also be represented as a set of attributes as suggested by Dolnicar and Otter (2003) and others (e.g., Qu et al., 2000). These attributes include services, location, room, price/value, food and beverage, image, security, and marketing. The frequently cited Two Factor Theory postulates that hygiene factors like cleanliness and maintenance do not positively contribute to satisfaction, although dissatisfaction results from their absence, while motivator factors such as the experiential aspects of staying at a hotel give positive satisfaction (Herzberg, 1966; Noe and Uysal, 1997). Recently, scholars have adopted the service-dominant logic arguing that guest experience is not be limited to what the hotel offers, but instead it is co-created by both the service provider and the hotel guest (Chathoth et al., 2013). Thus, guest satisfaction can be seen as the guest's evaluation of his/her experience through interaction with various service areas.

Given the complexity of the guest experience, measuring and managing hotel guest satisfaction is a challenging task. In the hospitality industry research has shown that there is a gap between what managers believe is important and what guests say is important in the selection and evaluation of accommodation (Lockyer, 2005). Consumer surveys, especially guest comment cards, have been widely used to measure hotel guest satisfaction (Pizam et al., 1982). Although it is efficient and useful, this method often suffers from poor sample quality and low response rates and produces generally vague assessments of a guest experience. Also, this type of survey does not take into account the importance of each of the hotel's individual attributes to the guest. Other measurement of hotel guest satisfaction such as importance–performance analysis can mitigate this type of problem. However, this approach requires that the hotel attributes being evaluated must be predefined. The use of open-ended questions, on the other hand, can generate rich and personally meaningful responses; however, the qualitative nature of such responses can be cumbersome to analyze and the results often lack generalizability (Crotts et al., 2009). At the conceptual level, in the attempt to measure guest satisfaction the validity of expectation measures associated with the expectancy-disconfirmation theory has been called into question. For example, there are different types of guest expectations and their relationships with other constructs in the satisfaction model can vary significantly, leading to unreliable outcomes (Oh, 1999).

As suggested by Oh (1999), it is important to consider new variables within the established conceptual framework in order to refine the theory about hotel guest satisfaction. While this statement was made from the conventional research perspective, to include and explore new data sources and novel analytical approaches to better understand guest experience and satisfaction seems to be a promising direction of research. In fact, there is a growing effort in using consumer-generated content to gauge guest/tourist satisfaction. For example, Pan et al. (2007) examined the usefulness of online travel blogs as a source of qualitative data describing guests' likes and dislikes in their purchase experiences. Crotts et al. (2009) applied a quantitative stance-shift analysis to measure hotel guest satisfaction using Internet blog narratives posted by guests. While these studies make valuable contributions to enrich our understanding of guest satisfaction, they relied upon a relatively small sample of online data to make inferences and therefore they are limited from the big data analytics standpoint. That is, although these studies may have high levels of internal validity, they may suffer to some extent from external validity issues since it would be difficult to generalize their findings on the basis of relatively small samples compared to large data sets.

2.3. Research questions

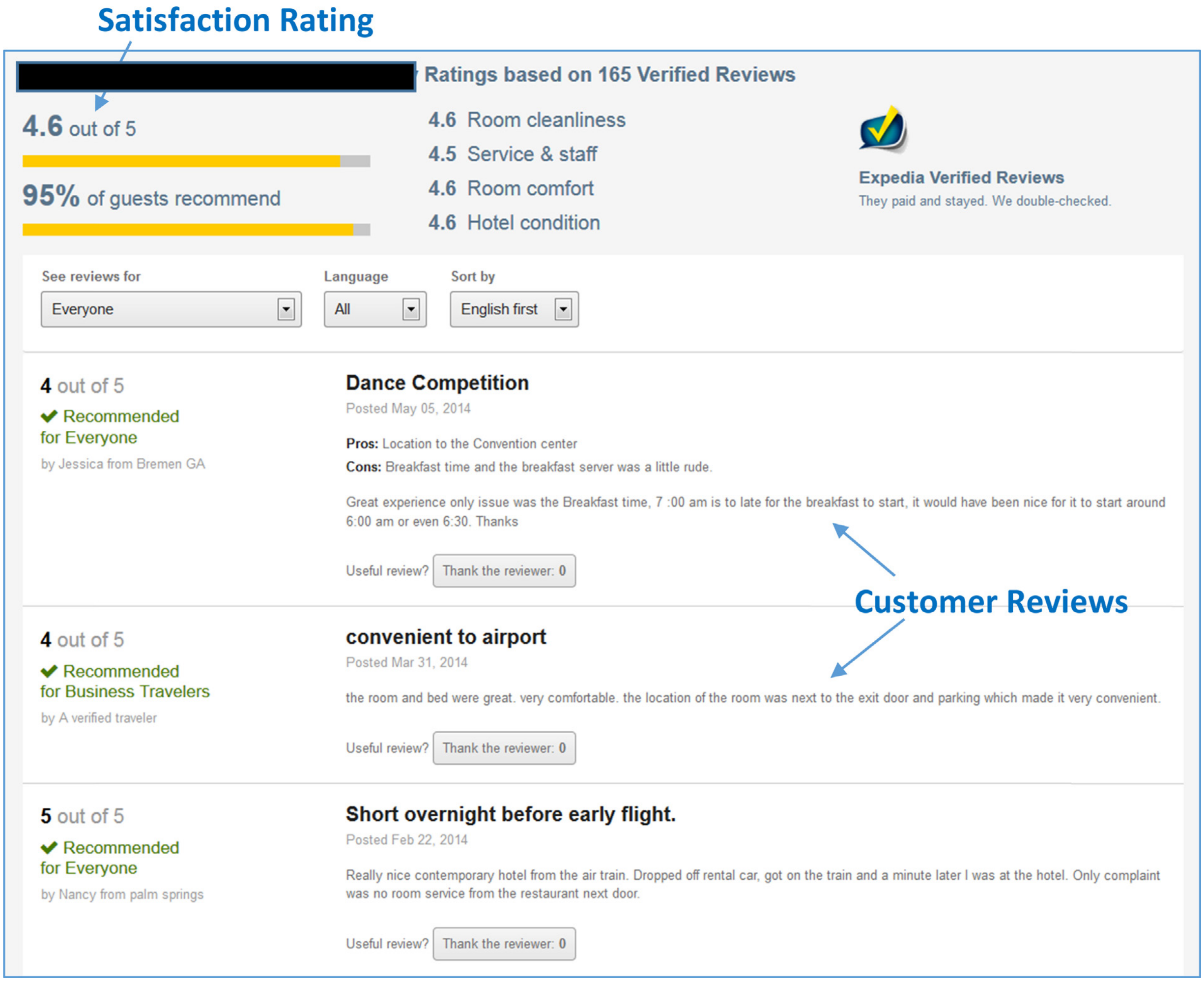
Online customer reviews have been widely considered one of the most influential types of consumer-generated content for understanding consumer behavior and consequently firm performance in hospitality and tourism (Browning et al., 2013; Serra Cantallops and Salvi, 2014; Mauri and Minazzi, 2013; Sparks and Browning, 2011). In many websites including TripAdvisor.com, and online travel agencies (OTAs) such as Expedia and Travelocity, consumers are allowed to post their ratings and reviews regarding their experiences with hotel properties they have stayed at in the past. Customer reviews reflect the way consumers describe, relive, reconstruct, and share their experiences. Because other consumers are tapping into this information for travel planning purposes, customer reviews can generate a huge impact on travel planning and subsequently attitudes and behavioral intentions (Gretzel and Yoo, 2008; Min et al., 2014). Importantly, the number of customer reviews has grown tremendously in recent years. For example, TripAdvisor claims that as of late 2013 there were more than 150 million reviews and opinions generated on its website alone covering more than 3.7 million accommodations, restaurants and attractions worldwide (see http://www.tripadvisor.com/PressCenter-c6-About_Us.html). In late 2012 Expedia's collection of verified reviews reached a total number of more than 7.5 million (see http://mediaroom.expedia.com). This wealth of consumer-generated data offers opportunities to describe, and make statistical inferences about, consumer behavior in hospitality. Following from above discussion, the following research questions were formulated to guide the study:

1. What is the nature and underlying structure of the hotel guest experience represented in customer reviews?
2. Can hotel guest experience represented in customer reviews be used to explain guest satisfaction?

3. Methodology

3.1. Research design

A large-scale text analytics study was conducted with the goal to understand hotel guest experience represented in online customer reviews and its association with satisfaction ratings based upon publicly available data in Expedia.com. Expedia.com was chosen because it is the largest online travel agency in the world with more than 16.5 million monthly unique visitors (see www.advertising.expedia.com). Also, unlike other websites that host consumer reviews, Expedia requires reviewers to make at least one transaction through its website before being allowed to contribute a review to the website. This essentially prevents hospitality businesses or marketers to post inauthentic reviews (Mayzlin et al., 2012). Usually after staying at the hotel property purchased through Expedia.com, the customer receives an email from the website soliciting feedback including ratings as well as his/her experience at the hotel. Fig. 1 displays a screenshot of a customer review page taken from Expedia.com. As can be seen, one hotel (name covered for anonymity) can have multiple entries of customer reviews. Along with these reviews there are also other types of information including hotel price, star rating, and customers' ratings. Customers assign an overall satisfaction rating with the property (4.5 out of 5 in this case), along with four additional ratings of room cleanliness, service and staff, room comfort, and



Satisfaction Rating

[redacted] Ratings based on 165 Verified Reviews

4.6 out of 5

95% of guests recommend

4.6 Room cleanliness
4.5 Service & staff
4.6 Room comfort
4.6 Hotel condition

Expedia Verified Reviews
They paid and stayed. We double-checked.

See reviews for: Everyone
Language: All
Sort by: English first

4 out of 5
✓ Recommended for Everyone
by Jessica from Bremen GA

Dance Competition
Posted May 05, 2014

Pros: Location to the Convention center
Cons: Breakfast time and the breakfast server was a little rude.

Great experience only issue was the Breakfast time, 7 :00 am is to late for the breakfast to start, it would have been nice for it to start around 6:00 am or even 6:30. Thanks

Useful review? Thank the reviewer: 0

Customer Reviews

4 out of 5
✓ Recommended for Business Travelers
by A verified traveler

convenient to airport
Posted Mar 31, 2014

the room and bed were great. very comfortable. the location of the room was next to the exit door and parking which made it very convenient.

Useful review? Thank the reviewer: 0

5 out of 5
✓ Recommended for Everyone
by Nancy from palm springs

Short overnight before early flight.
Posted Feb 22, 2014

Really nice contemporary hotel from the air train. Dropped off rental car, got on the train and a minute later I was at the hotel. Only complaint was no room service from the restaurant next door.

Useful review? Thank the reviewer: 0

Fig. 1. Screenshot of customer reviews on Expedia.com.

hotel condition. The main goal of the analysis was to understand the content and structure of customer reviews and how they are associated with hotel guest satisfaction (i.e., overall rating).

3.2. Data collection

Data were collected during the period of December 18–29, 2007 using an automated Web crawler (refer to Stringam and Gerdes (2010) for details). In a nutshell, the Web crawler visited Expedia.com and extracted customer reviews for all hotels listed by Expedia for the 100 largest U.S. cities, as defined then by the most recent U.S. Census Bureau population estimate (US Census Bureau, 2007). For each city the crawler gathered all available customer reviews and associated ratings for each hotel, adhering to the site's Robot Exclusion Standard restrictions (see http://en.wikipedia.org/wiki/Robots_exclusion_standard). The crawler collected data on 10,537 hotels resulting in 60,648 customer reviews, which means each hotel on average had approximately six customer reviews. Considering there were in total less than 50,000 hotel properties in the US (see American Hotel & Lodging Association's website http://www.ahla.com/), this represents more than one-fifth of the entire hotel population nationwide.

Once the data were collected, the extraction process identified all unique words contained in the text comments resulting in 6642 words from all customer reviews. This word bank, with frequencies ranging from words such as "hotel" (33,549) and "room" (22,213) to many words with a frequency of one, serves as the basis for understanding the domain of guest experience. Fig. 2 shows the distribution of these 6642 words based upon frequency. As can be seen, the distribution is highly skewed in that there are relatively a very small number of words with high frequencies while approximately three fifths of them have a frequency of one. A relational database was created using Microsoft Access with unique identifiers assigned to every hotel property, every customer review, and every unique word so that associations could be easily established for analytical purposes. For example, each hotel could be associated

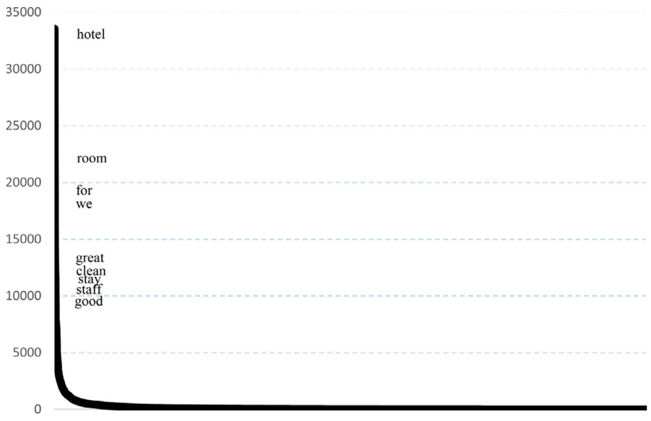


Fig. 2. Distribution of the Word Bank based upon frequency.

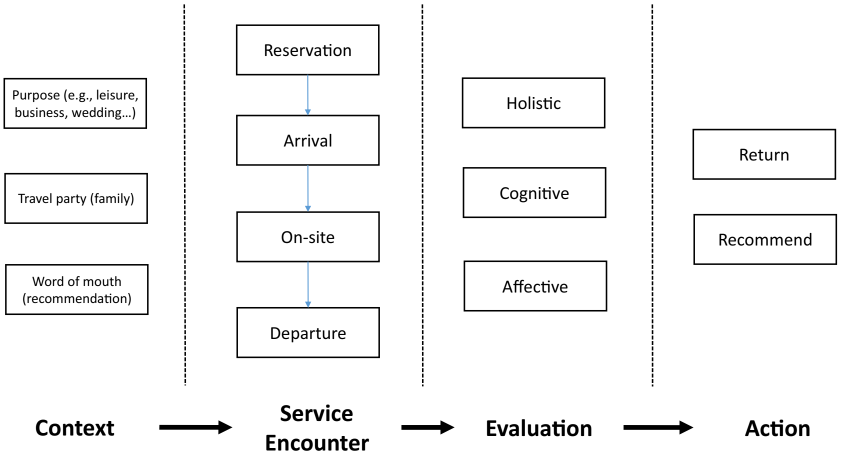


Fig. 3. Coding schema for hotel guest experience.

with a number of customer reviews which, in turn, were associated with multiple uniquely identified words. In total, this database contains about 1.3 million word-review pairs, which suggests that on average one customer review contains about 22 unique words (counting each word only once regardless of how many times it occurred in a specific review).

3.3. Data analysis

Data analysis followed a text analytics process which typically involves several steps including data pre-processing, domain identification/classification, and statistical association analysis (see Fan et al., 2006). While statistical analysis aims to examine the associations between the identified domain-related words and the dependent variable (i.e., hotel guest satisfaction in this case), the first two steps, i.e., data pre-processing and domain identification, are critical for establishing content validity with the focus on extracting conceptually relevant linguistic entities (words) from the corpus (Krippendorff, 2012). Typical textual data pre-processing involves a series of operations such as stemming (i.e., coding several forms of a linguistic entity into a 'rudimentary' form which represents the same meaning), misspelling identification, and identification and removal of stop words such as certain pronouns, adverbs, and conjunctions. Domain identification aims to classify guest experience-related words and non-guest experience-related ones. Normally, data pre-processing and domain identification are conducted in separate steps because they serve distinct purposes. However, since to our knowledge there was no readily available "dictionary" that describes hotel guest experience, these operations were conducted manually and simultaneously through an iterative process. Considering the sheer size of the word bank, this was a tedious and labor-intensive process. For example, there were a large number of variations for a word like "restaurant" with its different forms (e.g., single and plural) and many misspellings.

3.4. Data pre-processing and domain identification

To ensure validity and reliability a coding schema was established to guide the domain identification process in order to extract words related to hotel guest experience (see Fig. 3). This coding schema took into account the existing literature on each stage of the guest's experience with hotels services, i.e., reservation, arrival, on-site experience, and departure (Kotler et al., 2006). This was different from the commonly used sentiment analysis approach, which primarily aims to detect subjective evaluation of a product. In addition, compared to previous studies which solely focused upon salient attributes of a hotel (e.g., Crotts et al., 2009; Saleh and Ryan, 1992), this study took a broader, more comprehensive perspective to include (1) the context of guest experience, such as purpose of stay (business, leisure travel, and wedding, etc.), travel party (e.g., family), and reason for the purchase (e.g., due to friends' recommendation); (2) guests' verbal evaluation of the experience (e.g., "the staff was very friendly"); and, (3) expression of possible future actions (e.g., "I will recommend this hotel to my friends" or "I will return in the future"). It is believed that a broader perspective would have a better chance to capture the complex, idiosyncratic nature of personal experience and thus increase the likelihood for it to be related to the guest's overall satisfaction with a hotel property.

Coding was primarily conducted by one of the researchers and another researcher independently verified the coding results in terms of relevancy to the coding schema. Basically, any words that reflected those aspects of guest experience were included as part of the dictionary with a few exceptions: (1) stop words as mentioned above; (2) generic nouns such as "size", "people", "effort", and "fault", etc. due to the lack of specificity; (3) generic verbs such as "need", "want", "like", and "offer" because it was assumed meanings of these words were already captured in the objects of these verbs; (4) words with high ambiguity such as "break", "firm", "look", "ground", and "line" etc.; and, (5) words related to hotel brands such as "hilton", "marriott", and "ramada" since hotel identity was contained in the original downloaded dataset anyway. This coding process was done iteratively from words with high frequencies to those with low frequencies. The word "hotel" was not included in the dictionary because by default all these reviews were about hotel experience and, therefore, it was considered redundant. As part of the coding process, all possible variations of a specific word (e.g., plurals and misspellings) were manually searched and identified.

Coding stopped at the point when words encountered in the words bank appeared generic and irrelevant to the coding schema, resulting in a dictionary of 416 "primary" words that were used by consumers to describe their experiences at a specific hotel. Then, all the variations of these 416 words, when applicable, were substituted with each of the corresponding primary words. In total these 416 primary words and their variations represented roughly 40% (414,833/1,048,575) of occurrences among all 6642 unique words in the words bank. Frequencies of these 416 words were calculated for each of the 10,537 hotel properties using the PivotTable function in Microsoft Excel. This yielded a table of 5990 hotel cases because some of the customer reviews and thus the hotel cases did not contain any of these 416 words and were thus dropped. It is noteworthy that a word was counted only once regardless how many times it occurred in one customer review. Then, this frequency table was merged with the original dataset of hotel property attributes that includes location, satisfaction ratings, other ratings, as well as comments IDs each of these hotels was associated with. This master table was then imported into SPSS for statistical association analysis.

3.5. Statistical analysis

George et al. (2014) suggested that when dealing with big data, the focus of statistical analysis should be on effect sizes and variance explained instead of the conventional p value of relationships. Also, big data analytics does not necessarily mean more data is always better (Boyd and Crawford, 2012). With this in mind, the analysis focused on identifying guest experience-related words with the highest explanatory power on guest's satisfaction rating. This was achieved in an exploratory fashion from two "angles". First, the dataset was quite sparse for some hotel cases, thus, hotels with very low frequencies of dictionary words were removed. Second, many of these 416 words also had relatively low total frequencies. In fact, more than 80% of these words had occurred in less than 2% of all customer reviews. This, obviously, was a problem for our analysis when one of the assumptions was the covariance between variables (word frequencies). As such, words with very low frequencies were removed. This model was optimized by adjusting the hotel and word frequency thresholds to maximize the explanatory power on satisfaction rating in a linear regression. As a result, the dataset was reduced to 529 hotels and 80 guest experience-related words.

To answer the first research question, factor analysis was conducted to identify the underlying structure of customer reviews. Although the variables were non-metric, factor analysis was considered appropriate since the variables in focus are correlated to each other (Hair et al., 2009). More importantly, factor analysis extracts the communalities among these words, which, in this case, represented the connectivity of words in a specific factor because the variance is based upon co-occurrences of these words within the same customer reviews. In this way, extracted factors actually represented the common semantic space, i.e., the contexts in which words occurred. To answer the second research question, linear regression analysis was used to examine the relationship between guest experience and satisfaction using the factor scores as independent variables and average satisfaction rating as the dependent variable.

4. Findings

This section presents the main findings, in two parts: the first part provides a basic description of the final, "clean" data (with 529 hotel cases and 80 guest experience-related words) while the second part focuses on the statistical analysis that aimed to answer the research questions. Table 1 shows the distribution of the hotel properties across the US. As can be seen, these hotels were located in 32 states. Compared with the original dataset that contains the 5990 hotel properties with the 416 dictionary words, only hotels in the state of Alabama were removed during the cleaning process. A further examination showed that the number of cities where these hotels were located was reduced from the original 100 to 66. In the final dataset the top 10 cities with most hotels included New York, San Francisco, Orlando, San Diego, Seattle, Chicago, Boston, Miami, Portland, and Anaheim (in that order in terms of number of hotel properties). Interestingly, the top ten cities in the clean dataset had nearly 60% of the total number of properties while the top 10 in the original dataset contained only 34% of all hotel properties. This seems to suggest that hotel properties in the top 10 cities in the clean dataset tended to attract more customer reviews, at least in the case of Expedia.com. Also, this clearly shows that hotel properties were not treated equally in the social space and the distribution may reflect the volume of tourism as well as the overall hospitality atmosphere in these cities.

Table 1
Distribution of hotel properties across US used in analysis.

State	*N*	Percent (%)	Cumulative percent (%)
CA	101	19.1	19.1
FL	76	14.4	33.5
NY	58	11.0	44.5
WA	33	6.2	50.7
IL	29	5.5	56.2
TX	26	4.9	61.1
MA	24	4.5	65.6
OR	18	3.4	69.0
AZ	16	3.0	72.0
CO	16	3.0	75.0
LA	15	2.8	77.8
HI	14	2.6	80.4
VA	14	2.6	83.0
GA	11	2.1	85.1
PA	10	1.9	87.0
NV	9	1.7	88.7
MD	8	1.5	90.2
MN	7	1.3	91.5
NM	7	1.3	92.8
OH	5	.9	93.7
NC	4	.8	94.5
AK	3	.6	95.1
KY	3	.6	95.7
MI	3	.6	96.3
MO	3	.6	96.9
NE	3	.6	97.5
NJ	3	.6	98.1
TN	3	.6	98.7
WI	3	.6	99.3
IN	2	.4	99.7
KS	1	.2	99.9
OK	1	.2	100.0
All	529	100	

Table 2 shows the distribution of hotel properties based upon star rating provided by Expedia. As can be seen, hotels with star rating from two to four constituted more than 96% of all hotels. This is comparable to the original dataset wherein hotels with star rating between two and four constituted almost 98% of all hotels. An observable difference between these two datasets is that the clean dataset contained a higher percentage of hotels with a higher star rating (three and above). This seems to indicate that, while there were apparently more mid-range hotel properties than basic hostels/dormitories (below two stars) or luxury hotels (above four stars), customer reviews in Expedia.com tended to gravitate toward the mid-range hotels. Average satisfaction rating of hotels in the clean dataset is 4.02/5.0 with a standard deviation of .51, which is similar to the original dataset with a mean of 3.92/5.0 with a

Table 2
Distribution of hotel properties used in analysis based upon star rating.

Expedia star rating	Frequency	Percent (%)	Cumulative percent (%)
1.0	1	.2	.2
1.5	14	2.6	2.8
2.0	93	17.6	20.4
2.5	78	14.7	35.2
3.0	163	30.8	66.0
3.5	111	21.0	87.0
4.0	65	12.3	99.2
4.5	3	.6	99.8
5.0	1	.2	100.0
Total	529	100.0	

Table 3
Top 80 primary words in hotel customer reviews.

Word	*N*	*N*/Hotel	Word	*N*	*N*/Hotel	Word	*N*	*N*/Hotel	Word	*N*	*N*/Hotel
Room	5641	10.7	Downtown	676	1.3	Lobby	357	.7	Experience	240	.5
Clean	3104	5.9	Airport	620	1.2	Internet	344	.7	Suite	236	.4
Staff	2898	5.5	Desk	609	1.2	Trip	328	.6	Money	233	.4
Location	2865	5.4	View	569	1.1	Pay	320	.6	Carpet	233	.4
Comfortable	2168	4.1	Recommend	532	1.0	Door	317	.6	Courteous	233	.4
Service	1707	3.2	Noise	493	.9	Shops	316	.6	City	231	.4
Friendly	1614	3.1	Quiet	486	.9	Sleep	303	.6	Expensive	223	.4
Close	1594	3.0	Food	468	.9	Business	301	.6	Dirty	221	.4
Breakfast	1524	2.9	Distance	464	.9	Complaint	299	.6	Renovated	219	.4
Helpful	1378	2.6	Shuttle	447	.8	Shower	296	.6	Tub	217	.4
Bed	1334	2.5	Street	429	.8	Family	294	.6	Safe	216	.4
Price	1321	2.5	Shopping	419	.8	Value	290	.5	Far	214	.4
Restaurants	1153	2.2	Maintained	417	.8	Cheap	288	.5	Air	213	.4
Walking	1011	1.9	Beach	398	.8	Smelled	284	.5	Refrigerator	205	.4
Area	863	1.6	Access	398	.8	Kids	258	.5	Quality	203	.4
Parking	802	1.5	Park	385	.7	Tv	256	.5	Decor	201	.4
Bathroom	764	1.4	Floor	373	.7	Attractions	248	.5	Wait	200	.4
Pool	716	1.4	Check in	369	.7	Water	247	.5	Freeway	198	.4
Free	712	1.3	Spacious	365	.7	Coffee	244	.5	Elevator	196	.4
Convenient	708	1.3	Bar	358	.7	Amenities	244	.5	Accommodation	114	.2

standard deviation of .74. In the clean dataset, average number of reviews per hotel is greater than in the original dataset (approximately 48 vs. 15). This shows, from a different angle, the extent to which the sparseness in the original dataset was reduced.

Table 3 provides the list of the 80 guest experience-related words that were used to explain satisfaction ratings along with their total frequency and average frequency per hotel. These words reflect a wide spectrum of aspects related to the hotel guest experience, including (1) the very core product such as “room”, “bed”, and “bathroom”; (2) hotel amenities such as “front” (desk), “restaurant”, “pool”, “parking”, “lobby”, “shower”, “TV”, “bar”, and “amenities”, etc.; (3) hotel attributes such as “location”, “downtown”, “close”, service”, “price”, “walking”, “distance”, “airport”, “free”, “view”, “quiet”, “noise”, “far”, “renovated”, and others; (4) hotel staff-related descriptors such as “staff”, “friendly”, “helpful”, and “courteous”; (5) hotel service encounters such as “parking”, “checkin”, “shopping”, “complaint”, “wait”, and “pay”; (6) evaluation of experience such as “clean”, “comfortable”, “maintained”, “safe”, “smelled”, “value”, and “cheap”; (7) travel context such as “business” and travel party such as “family”, “kids”, and “husband”; and, (8) possible actions such as “recommend”. Compared with the coding schema, this list does not reflect certain aspects of guest experience such as stay at the hotel due to word-of-mouth (recommendations), the departure stage (checkout) of service encounters, affective evaluation of the experience, as well as other possible actions after the stay, etc.

The frequency distribution of these 80 words is highly skewed, in that the top 12 words constitute more than half, and the top 25 words nearly 70%, of the total frequency of all words. This distribution can be characterized as one with a “head”, i.e., word with relatively high frequencies, and a “long tail”, i.e., those with low frequencies (with an average frequency per hotel of less than 1 starting from the 26th word). The “head” words center around the core and basic products/services as well as important attributes such as the guest room, cleanliness, staff, location, comfort, service, friendliness and helpfulness of staff, breakfast, bed, and price, etc. The “long tail” words reflect other important areas of guest experience. Generally speaking, most of these words are functional and objective, while a handful of them represent guests’ subjective evaluation of their hotel experience. It is interesting to note that words denoting travel party (‘family’ in this case), food-related aspects such as breakfast, restaurants, bar, and even coffee, and activities guests can do outside of the hotel property such as shopping and visit to the beach, are also relevant to guest experience. Overall these 80 words reflect a diverse array of amenities, attributes, and service encounters shaped by hotel guests’ unique expectations and evaluations at the aggregate level.

Factor analysis was employed in order to examine the underlying semantic structure and further reduce the number of words from the data matrix into meaningful groupings of words that would be easier to interpret. As can be seen in Table 4, six meaningful factors consisting of 34 words out of the final 80 words in Table 3 emerged from the factor analysis explaining 22.84% of all variance. Keep in mind that, different from factor analysis based upon metric data, factors obtained from this analysis represent the common semantic spaces in customer reviews. Since the loadings were relatively low (compared to factor analysis conducted using established metric scales), the cutoff loading was set at (±).30 in order to capture as many words as possible. Also, the cutoff eigenvalue was set at 2 because, as the number of factors increases, the more difficult it becomes to interpret those “small” factors. Each factor was named based upon the semantic space represented by the words in the specific factor. The first factor, containing 14 words, was named “Hybrid” because it appears to be comprised of two distinctive groups of words that represent very different hotel guest experiences. The first group of words, including “clean”, “smelled”, “dirty”, “price”, “cheap”, “carpet”, and “sleep”, seems to be dominated by maintenance-related aspects which could affect the guest’s basic needs (“sleep”) and perception of product (“cheap”). The second group of words, including “expensive”, “shopping”, “view”, “restaurants”, “distance”, “location”, and “walking”, seems to represent the experiential aspects of the hotel stay, particularly in words such as “shopping”, “restaurant”, “location”, “walking”, and “view”. What is revealing is that these two groups of words have the opposite signs in their loadings: loadings in the first group are all positive while in the second group all negative. This suggests that, in the semantic space that represents hotel guest experience, these two groups of words belong to two very different contexts of meaning. That is, when a consumer mentions the words in the first group, he/she is unlikely to use words in the second group to describe the experience. Behaviorally speaking, it seems the maintenance-related aspects are “blocking” the experiential aspects of the hotel stay in the guest’s mental model. In other words, the presence of any maintenance factors associated with “smelled”, “dirty”, “price”, “cheap”, “carpet”, and “sleep” may not add much to satisfaction but their absence will certainly detract from satisfaction.

The other five factors are quite straightforward to interpret. Factor 2 was named “Deals” apparently because the word “free” occurred with “breakfast”, “airport”, and “shuttle”. The third factor

Table 4
Factor loadings of words (shows only those with loadings > 30).

Words (N = 34)	Factor loadings					
	F1	F2	F3	F4	F5	F6
Hybrid						
Clean (5.9)[a]	.436					
Smelled (.5)	.423					
Dirty (.4)	.395					
Price (2.5)	.369					
Cheap (.5)	.354					
Carpet (.4)	.349					
Sleep (.6)	.323					
Expensive (.4)	−.313					
Shopping (.8)	−.326					
View (1.1)	−.377					
Restaurants (2.2)	−.387					
Distance (.9)	−.459					
Location (5.4)	−.492					
Walking (1.9)	−.496					
Deals						
Breakfast (2.9)		.517				
Airport (1.2)		.443				
Free (1.3)		.435				
Comfortable (4.1)		.409				
Shuttle (.8)		.393				
Amenities						
Close (3.0)			.390			
Beach (.8)			−.366			
Pool (1.4)			−.533			
Family friendliness						
Family (.6)				.509		
Kids (.5)				.483		
Attractions (.5)				.338		
Suite (.4)				.313		
Service (3.2)				−.338		
Core product						
Room (10.7)					.552	
Bathroom (1.4)					.420	
Bed (2.5)					.322	
Spacious (.7)					.302	
Staff						
Helpful (2.6)						−.462
Friendly (3.1)						−.511
Staff (5.5)						−.517
Eigenvalue	4.55	3.66	3.07	2.66	2.30	2.05
Cumulative variance	5.69%	10.26%	14.09%	17.41%	20.28%	22.84%

[a] Indicating average number of times this word occurred in a hotel's customer reviews (based upon Table 3).

"Amenities" consists of only three words, with "beach" and "pool" having a negative sign suggesting that when customers mention the word "close", it is unlikely referring to "beach" and "pool". This implies that these two words tend to have a negative connotation when customers talk about convenience and access to amenities. The fourth factor, i.e., "Family Friendliness", seems to suggest that, when customers share their story about staying at a hotel with their family members, their experience is likely to be linked with the need for a large room ("suite") or attractions they want to visit. It is unlikely for them to talk or care about the hotel service. The fifth factor reflects the core product of a hotel, i.e., the guest room, bed, and bathroom. It is interesting to note the word "spacious" is used within this context. Lastly, the sixth factor represents customers' perception of hotel staff with words such as "helpful" and "friendly". All three words have negative loadings on this factor, suggesting that, in general, there is a negative connotation to the context wherein customers mentioned their experience with hotel staff.

Overall these factors captured the salient aspects of hotel guest experience in that most of the primary words with high frequencies in customer reviews generated relatively high loadings on these factors. Some long tail word such as "shopping", "distance", "beach", "spacious", "sleep", "family", "kids", "smelled", "attractions", "suite", and "expensive", also contributed to these factors. While some factors such as travel party (i.e., family in this case) seemed to be highly relevant to guest experience, other factors traditionally considered important such as front desk services, did not have significant impact on the semantic space representing hotel guest experience based on the customer reviews.

Table 5 shows the ANOVA results using average satisfaction rating as the dependent variable and the six hotel guest experience factors as independent variables. All factors except Amenities were significant at the p = .01 level, with the first two factors, Hybrid and Deals having the largest standardized coefficients of −.567 and .506, respectively. This suggests that Hybrid and Deals are the most important factors associated with guest satisfaction. Interestingly, the factor Core Product, although significant, was not as important as Hybrid, Deals, and Family Friendliness. The signs of these coefficients are quite revealing: the negative sign for Hybrid suggests that this factor, represented by the 14 guest experience-related words, connotes a negative meaning for guest satisfaction. Since the factor loadings of the Hybrid maintenance and cleanliness-related words are positive while the factor loadings of the Hybrid experiential words are negative, this means that this factor carries a negative "sentiment": if satisfaction rating is low, hotels reviewed by Expedia customers tend to be NOT well maintained and did NOT support experience co-creation by the customer. In the case of factor Deals, since the coefficient is

Table 5
Results of linear regression analysis.

Model	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	Std. error	Beta		
(Constant)	4.023	.013		298.410	.000
Hybrid	−.293	.013	−.576	−21.714	.000
Deals	.258	.013	.506	19.086	.000
Amenities	−.015	.013	−.029	−1.076	.282
Family friendliness	.076	.013	.149	5.606	.000
Core Product	.063	.013	.123	4.641	.000
Staff	.044	.013	.086	3.242	.001

Dependent variable: average customer rating; Adjusted *R* square: .629.

positive, it means that a high satisfaction rating is associated with the mentions of words about free services (i.e., breakfast and airport shuttle). It is interesting in the case of Family Friendliness in that most of the words about aspects related to traveling family members have positive factor loadings, suggesting a higher satisfaction rating is associated with mentions of these words. However, the negative sign for the word "service" suggests that when a high satisfaction score is associated with the word "service" NOT being mentioned in the context of those words. Staff-related words are negatively loaded on to the factor Staff suggesting a high satisfaction rating is not likely associated with the mentions of words such as "helpful" and "friendly".

5. Discussion

Hotel guest experience and satisfaction have been extensively studied in the hospitality management literature. Guest experience is, undoubtedly, an extremely complex construct. Depending upon the research design and methods researchers could get very different pictures of what constitutes guest experience and what actually leads to guest satisfaction (see Crotts et al., 2009; Lockyer, 2005; Pizam et al., 1982). Since conventional methods usually rely on a set of predefined hypotheses, justified using previous and existing body of knowledge, the attempts are made in the direction of either confirming or disconfirming such hypotheses. However, this is not the case with big data analytics. Through the analytical process we as researchers let the data reveal patterns reflective of consumers' reliving and evaluation of their actual experiences with products (hotels in this case). Then, we attempted to make sense and attach meaning to the inferences by bringing appropriate theories to shed light on and explain revealed/novel patterns from large data. Different from conventional methods this way of explaining the findings is part of epistemology of generating and creating knowledge using big data (George et al., 2014). Although there is no previous study to benchmark against, the validity of our study, like many others based upon big data, was established by the meticulously devised analytical process that strictly followed both theory (e.g., content analysis and the definition of hotel product and guest experience) and common practices in text mining.

Compared to other text analytics approaches such as sentiment analysis, which generally aim to capture the subjective opinions of online consumers about certain products (Pang and Lee, 2008), this study is unique in its analytical process. First, this study set out with the goal to enrich our existing knowledge about a theoretical construct. In addition to the standard operations such as stemming and stop words identification normally used in text analytics, a conceptual framework was employed as a coding schema during the analytical process in order to capture, to the greatest extent possible, the domain of guest experience. As shown in this study and many others, big data can contain plenty of noise. The iterative analytical process, i.e., reducing data scarcity one at a time, was also critical for identifying a robust data structure that yielded strong, meaningful associations between two distinct domains of variables. Therefore, our approach reflects an exploratory process guided by theory. One drawback in our text analytical approach was that it did not apply word-sense disambiguation and semantic valence detection during the coding process, which could lead to the loss of variance in the data. However, this implies that the identified associations between variables could be even stronger had these techniques been applied to the analysis.

The dictionary identified for hotel guest experience reflects what consumers think are relevant and important that contribute to their (dis)satisfaction with a specific hotel (Stringam and Gerdes, 2010). As such, this list of words is a "discrete" representation of guest experience rank-ordered by word frequency. More importantly, the structure of guest experience identified through factor analysis is particularly revealing in that guest experience, to a great extent, can be represented by a handful of underlying dimensions that, although not completely different from existing literature, carry varying weights as well as have meaningful semantic compositions. Among these dimensions, of particular interest is the Hybrid factor which seems to be dominant and quite complex in its own. It was somewhat counterintuitive at first sight that two totally different or even mutually irrelevant groups of words were "lumped" together in the same dimension. However, with careful consideration of the semantic nature of customer reviews, this factor actually reveals an interesting aspect of customer reviews in that the use of one set of words (i.e., maintenance and cleanliness) appears to "block" the use of another (i.e., experiential aspects of the stay). That is, these words contributed to the same dimension but they were used in completely different contexts. Conceptually, at first sight these two sets of words appear to be in line with the frequently cited Two-Factor Theory of Motivation (Herzberg, 1966) and its variants in the field of hospitality and tourism (e.g., Noe and Uysal, 1997). These theories postulate that hygiene or instrumental factors like cleanliness and maintenance do not positively contribute to satisfaction, although dissatisfaction results from their absence, while motivators or expressive factors such as the experiential aspects of staying at a hotel give positive satisfaction. Importantly, according to these theories these two types of factors contribute to satisfaction independent of each other (e.g., the staff can be friendly regardless if the hotel is clean or dirty). In contrast, our analysis shows that these two types of factors are inherently connected to each other and the level of the experiential, co-produced satisfiers is highly dependent upon the hygiene factor level. This suggests, for example, if a guest is upset about the dirty room or lack of maintenance, it is very unlikely for him/her to be interested or fully engaged in activities that promote experiential encounters or co-creation of the experience (Chathoth et al., 2013).

In a similar way, a guest who stays with family members seems to be not interested in the service aspect of the hotel other than a spacious room and attractions nearby. This indicates that, within the consumer's complex mental model about the hotel experience, there are structures of "domains" that are mutually exclusive, or that one serves as the necessary condition for another. This also points to the fact that because of the tangible aspects of maintenance factors, hotels should provide and develop appropriate service amenities and features, and maintain them at the performance level that is expected to be in place. Another important insight is the identification of the Family Friendliness factor, which shows that what the guest brings into the experience, i.e., the travel party, can be an important contributing factor to their satisfaction. In addition, some of the "long tail" words in all of these dimensions show that the underlying semantic structures in customer reviews could be more conceptually relevant than simply words with high frequencies (Stringam and Gerdes, 2010; Xiang et al., 2009). These insights also attest to the capabilities of big data analytics to identify novel patterns through unconventional analytical approaches.

Although guest satisfaction is not measured in the traditional sense, the association between satisfaction rating and guest experience appears to be strong. According to Lewis (1985), when hotel guest satisfaction is being examined as the dependent variable, an R square value between .50 and .60 is considered acceptable. Our study showed that the underlying factors representing a set of only 34 words can explain nearly 63% of the total variance in guest satisfaction, which considerably exceeded the acceptable range. This indicates guest experience represented in customer reviews is highly associated with guest satisfaction; or, more precisely, it shows a general pattern that a customer tends to use particular words to describe his/her experience when he/she is happy or unhappy about the hotel. This is also quite different from conventional approaches, which solicit responses to a pre-established schema (Crotts et al., 2009), in that it shows these two domains of consumer behavior, i.e., experience and satisfaction, are inherently and "naturally" connected. Considering that guest satisfaction is measured as the average rating of all customers who reviewed the same hotel and the effect of these words in customer reviews could have been "evened out", this association could be even stronger had the analysis been done using cases of individual customers instead of cases of hotels. While the semantic compositions identified in this study are arguably data specific, the findings clearly show that text analytics using customer reviews has the potential to enrich our existing knowledge about hotel guest experience and satisfaction.

6. Conclusions and implications

While big data analytics has been touted as a new research paradigm in many disciplines, we have seen very few applications in the field of hospitality that fully explore its capabilities. This study applies text analytics to classify a large amount of online customer reviews, assess the quality of these data, as well as identify inherent relationships between two domains of variables in hotel management. The uniqueness of this study lies in the use of large data and delineation of guest experience drivers on a scale that was not available in traditional guest survey studies. Although this study is a preliminary effort in big data analytics, we have gained substantial insights into some of the extensively studied constructs in hospitality. As such, it is hoped that this study sets an example for the development of business analytics in hospitality marketing and management.

Our study contributes to the literature in several ways. First, this study demonstrated the usefulness of big data analytics in identifying novel patterns of hotel guest behavior using consumer generated content readily available on the Internet. While our findings were based upon data generated on a specific website and during a specific time period, they reflected the way consumers "talk about" their experiences in online reviews. As shown in the hybrid dimension of guest experience, our analysis revealed the semantic differentiations in relation to the hygiene and motivation factors expressed in online reviews. Also, the strong association between experience and satisfaction seems to be, to a great extent, a generalizable pattern. That is, it would be hard for us to imagine that the inherent connection between these two domains of consumer behavior would be substantially different over time or given another group of online consumers (e.g., Expedia vs. Travelocity). Or, at least, these findings can be seen as testable propositions derived from the analysis. This study has, once again, shown that there is always room for improvement in our knowledge about these well-studied phenomena in the hospitality field. Second, this study attests to the notion that theory should play a central role in big data analytics (George et al., 2014). Domain knowledge proved to be critical in guiding the data processing and analytical process before reaching the point where meaningful relationships emerged. Using effect size as the key indicator to reduce noise and improve data robustness was shown to be a sensible approach to understanding domain specific text data. Third, from a practical viewpoint this study offers food for thought for hotel management and marketing in that our analysis suggests hygiene factors are essential in hotel services without which the guest cannot function as a co-creator of the experience or at least cannot fully enjoy the experience. In addition, the strong association between satisfaction and guest experience has many potentially useful applications. For example, for service providers who do not directly monitor satisfaction (e.g., using guest comment cards), they might be able to infer levels of stratifications by "listening to" the words customers used to describe their experiences on channels such as Twitter and Facebook where satisfaction is generally not reported.

This study has several limitations and the findings should be interpreted with caution. Particularly, it is well known that there is self-selection bias when customers post online reviews (Li and Hitt, 2008). For example, customer satisfaction rating tends to be more on the positive side as clearly shown in our data. Another limitation was that the sample represented only urban hotels in the top 100 metropolitan areas in the US. The hotel attributes identified in customer reviews obviously reflected the perceptions of location-related aspects of the hotels. Guest experience could be considerably different in less populated, rural areas. Another limitation is that the data were collected several years ago and thus may not reflect current consumer sentiment. Nonetheless, the potential limitations in the generalizability of the findings does not reduce the internal validity of the data and thus does no harm to the purpose of demonstrating the power of big data analytics in the field of hospitality. Future research may consider applying methods of triangulation to multiple sources of data to validate the semantic structure of guest experience in order to develop a more comprehensive knowledge about guest satisfaction using big data analytics.

References

Abrahams, A.S., Jiao, J., Wang, G.A., Fan, W., 2012. Vehicle defect discovery from social media. Decis. Support Syst. 54 (1), 87–97.

Aiden, E., Michel, J.-B., 2014, April. The Predictive Power of Big Data. Newsweek, Available at http://www.newsweek.com/predictive-power-big-data-225125

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. J. Finance 59 (3), 1259–1294.

Boyd, D., Crawford, K., 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15 (5), 662–679.

Browning, V., So, K.K.F., Sparks, B., 2013. The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels. J. Travel Tour. Mark. 30, 23–40.

Chathoth, P., Altinay, L., Harrington, R.J., Okumus, F., Chan, E.S., 2013. Co-production versus co-creation: a process based continuum in the hotel service context. Int. J. Hosp. Manage. 32, 11–20.

Choi, T.Y., Chu, R., 2001. Determinants of hotel guests' satisfaction and repeat patronage in the Hong Kong hotel industry. Int. J. Hosp. Manage. 20 (3), 277–297.

Crotts, J.C., Mason, P.R., Davis, B., 2009. Measuring guest satisfaction and competitive position in the hospitality and tourism industry an application of stance-shift analysis to travel blog narratives. J. Travel Res. 48 (2), 139–151.

Dolnicar, S., Otter, T., 2003. Which Hotel Attributes Matter? A Review of Previous and a Framework for Future Research.

Duan, W., Gu, B., Whinston, A.B., 2008. Do online reviews matter? An empirical investigation of panel data. Decis. Support Syst. 45 (4), 1007–1016.

Engel, J.F., Blackwell, R.D., Miniard, P.W., 1990. Consumer Behavior, 6th ed. Dryden Press, Hinsdale, IL.

Fan, W., Wallace, L., Rich, S., Zhang, Z., 2006. Tapping the power of text mining. Commun. ACM 49 (9), 76–82.

Finch, B.J., 1999. Internet discussions as a source for consumer product customer involvement and quality information: an exploratory study. J. Oper. Manage. 17 (5), 535–556.

George, G., Haas, M.R., Pentland, A., 2014. Big data and management. Acad. Manage. J. 57 (2), 321–326.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.

Ghose, A., Ipeirotis, P.G., 2011. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. Knowledge and data engineering. IEEE Trans. Knowl. Data Eng. 23 (10), 1498–1512.

Ghose, A., Ipeirotis, P.G., Li, B., 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd sourced content. Mark. Sci. 31 (3), 493–520.

Gretzel, U., Yoo, K.H., 2008. Use and impact of online travel reviews. In: Information and Communication Technologies in Tourism 2008. Springer, Vienna, pp. 35–46.

Hair, J.F., Black, W.C., Barry, J., Babin, B.J., Anderson, R.E., 2009. Multivariate Data Analysis, 7th ed. Prentice Hall, Upper Saddle River, NJ.

Herzberg, F., 1966. Work and the Nature of Man. World Publishing, Cleveland.

Hunt, J.D., 1975. Image as a factor in tourism development. J. Travel Res. 13, 3–7.

Kotler, P., Bowen, J.T., Makens, J.C., 2006. Marketing for Hospitality and Tourism. Pearson Education, India.

Krippendorff, K., 2012. Content Analysis: An Introduction to its Methodology. Sage, Thousand Oaks, CA.

Lewis, R.C., 1985. Getting the most from marketing research: Part V. Predicting hotel choice: the factors underlying perception. Cornell Hotel Restaur. Adm. Q. 25 (4), 82–96.

Li, X., Hitt, L.M., 2008. Self-selection and information role of online product reviews. Inf. Syst. Res. 19 (4), 456–474.

Lockyer, T., 2005. The perceived importance of price as one hotel selection dimension. Tour. Manage. 26 (4), 529–537.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, Available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Marcus, G., Davis, E., 2014, April. Eight (no, nine!) problems with big data. The New York Times, Available at: http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Mattila, A.S., O'Neill, J.W., 2003. Relationships between hotel room pricing, occupancy, and guest satisfaction: a longitudinal case of a midscale hotel in the United States. J. Hosp. Tour. Res. 27 (3), 328–341.

Mauri, A.G., Minazzi, R., 2013. Web reviews influence on expectations and purchasing intentions of hotel potential customers. Int. J. Hosp. Manage. 34, 99–107.

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, New York, NY.

Mayzlin, D., Dover, Y., Chevalier, J.A., 2012. Promotional Reviews: An Empirical Investigation of Online Review Manipulation (No. w18340). National Bureau of Economic Research.

Min, H., Lim, Y., Magnini, V.P., in press. Responding to negative online hotel reviews: the impacts of empathy statements, paraphrasing and speed. Cornell Hosp. Q.

Noe, F.P., Uysal, M., 1997. Evaluation of outdoor recreational settings: a problem of measuring user satisfaction. J. Retail. Consum. Serv. 4 (4), 223–230.

Oh, H., 1999. Service quality, customer satisfaction, and customer value: a holistic perspective. Int. J. Hosp. Manage. 18 (1), 67–82.

Oh, H., Parks, S.C., 1997. Customer satisfaction and service quality: a critical review of the literature and research implications for the hospitality industry. Hosp. Res. J. 20, 35–64.

Oliver, R.L., 1981. Measurement and evaluation of satisfaction processes in retail settings. J. Retail. 57, 25–48.

Pan, B., MacLaurin, T., Crotts, J.C., 2007. Travel blogs and the implications for destination marketing. J. Travel Res. 46 (1), 35–45.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Found. Trends Inform. Retr. 2 (1/2), 1–135.

Pizam, A., Lewis, R.C., Manning, P., 1982. The Practice of Hospitality Management.

Qu, H., Ryan, B., Chu, R., 2000. The importance of hotel attributes in contributing to travelers' satisfaction in the Hong Kong Hotel Industry. J. Qual. Assur. Hosp. Tour. 1 (3), 65–83.

Saleh, F., Ryan, C., 1992. Client perceptions of hotels: a multi-attribute approach. Tour. Manage. 13 (2), 163–168.

Schmunk, S., Höpken, W., Fuchs, M., Lexhagen, M., 2013. Sentiment analysis: extracting decision-relevant knowledge from UGC. In: Xiang, Z., Tussyadiah, I. (Eds.), Information and Communication Technologies in Tourism 2014. Springer International Publishing, New York, NY, pp. 253–265.

Schumaker, R.P., Chen, H., 2009. Textual analysis of stock market prediction using breaking financial news: the AZF in text system. ACM Trans. Inform. Syst. 27 (2), 1–19.

Serra Cantallops, A., Salvi, F., 2014. New consumer behavior: a review of research on eWOM and hotels. Int. J. Hosp. Manage. 36, 41–51.

Sparks, B.A., Browning, V., 2011. The impact of online reviews on hotel booking intentions and perception of trust. Tour. Manage. 32 (6), 1310–1323.

Stringam, B.B., Gerdes Jr., J., 2010. An analysis of word-of-mouse ratings and guest comments of online hotel distribution sites. J. Hosp. Mark. Manage. 19 (7), 773–796.

Su, A.Y.L., 2004. Customer satisfaction measurement practice in Taiwan hotels. Int. J. Hosp. Manage. 23 (4), 397–408.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. J. Financ. 63 (3), 1437–1467.

US Census Bureau, Population Division, 2007, June. Table 1: Annual Estimates of the Population for Incorporated Places Over 100,000, Ranked by July 1, 2006 Population: April 1, 2000 to July 1, 2006 (CSV), Retrieved from: http://www.census.gov/popest/states/NST-ann-est2006.html

Wood, S.A., Guerry, A.D., Silver, J.M., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. Sci. Rep. 3, 2976.

Wu, C.H.J., Liang, R.D., 2009. Effect of experiential value on customer satisfaction with service encounters in luxury-hotel restaurants. Int. J. Hosp. Manage. 28 (4), 586–593.

Xiang, Z., Gretzel, U., 2010. Role of social media in online travel information search. Tour. Manage. 31 (2), 179–188.

Xiang, Z., Gretzel, U., Fesenmaier, D.R., 2009. Semantic representation of tourism on the Internet. J. Travel Res. 47 (4), 440–453.

Yang, Y., Pan, B., Song, H., 2013. Predicting hotel demand using destination marketing organization's web traffic data. J. Travel Res., http://dx.doi.org/10.1177/0047287513500391.

Ye, Q., Law, R., Gu, B., 2009a. The impact of online user reviews on hotel room sales. Int. J. Hosp. Manage. 28 (1), 180–182.

Yoon, Y., Uysal, M., 2005. An examination of the effects of motivation and satisfaction on destination loyalty: a structural model. Tour. Manage. 26 (1), 45–56.