Open Access Library Journal
2018, Volume 5, e4877
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Exploring Big Data Applied in the Hotel Guest Experience

Chieh-Heng Ko

Department of Hospitality Management, College of Tourism and Hospitality, Da-Yeh University, Taiwan
Email: chko@mail.dyu.edu.tw

How to cite this paper: Ko, C.-H. (2018) Exploring Big Data Applied in the Hotel Guest Experience. *Open Access Library Journal*, **5**: e4877.
https://doi.org/10.4236/oalib.1104877

Received: August 31, 2018
Accepted: October 12, 2018
Published: October 15, 2018

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.

Subject Areas

Economics, Marketing

Keywords

Big Data, Text Analytics, Guest Experience, Hotel Management

1. Introduction

Social media and consumer-generated content on the Internet continue to grow

and impact the hospitality industry [1]. 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 [2]. The so-called big data analytics approach emphasizes and lever-ages the capacity to collect and analyze data with an unprecedented breadth, depth, and scale to solve real-life problems. 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 [3]. 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. 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 [4].

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 relation-ships 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 collec-

tion 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. Literature Review

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 [2]. 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 [5]. 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 [6]. 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 [6].

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 [7]. Mining social media and consumer-generated content has attracted much attention for their value as public and community data. For instance, research has demonstrated that online consumer reviews can be used to predict product quality (Finch, 1999), stock market volatility and box office sales in the movie industry [8]. It has been found that online news postings have sufficient linguistic content to be predictive of a firm's earnings and stock returns [9]. More recently, Ghose and Ipeirotis [10] used text content and reviewer characteristics to estimate the helpfulness and economic impact of online hotel product reviews. Abrahams *et al.* [11] 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 [12].

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 [13]. 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 [14]. 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 [15] 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 [16]) define customer satisfaction as an emotional response to the use of a product or service. Oh [16] postulate that satisfaction involves cognitive and affective processes, as well as other psycho-logical 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 [17]. 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 [18].

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. The hotel product can also be represented as a set of attributes as suggested by Dolnicar and Otter [19]. 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 con-tribute to satisfaction, although dissatisfaction results from their absence, while motivator factors such as the experiential aspects of staying at a hotel give positive satisfaction. 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 [20]. 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 [21]. Consumer surveys, especially guest comment cards, have been widely used to measure hotel guest satisfaction. 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 pre-defined. 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 [22]. 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 [23].

As suggested by Oh [23], 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.* [24] 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.* [22] applied a quantitative stance-shift analysis to measure hotel guest satis-

faction 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 [25]. 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 [26]. 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-AboutUs.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. 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.

3.2. Data Collection

Data were collected during the period of December 18-29, 2017 using an automated Web crawler. In a nutshell, the Web crawler visited Expedia.com and extracted customer reviews for all hotels listed by Expedia in Taiwan. The crawler collected data on 106 hotels resulting in 6027 customer reviews, which means each hotel on average had approximately sixty customer reviews. 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. 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 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. 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. 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.

4. Results

Table 1 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", "down-town", "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", "check-in", "shopping", "complaint", "wait", and "pay"; 6) evaluation of experience such as

Table 1. 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	0.7	Experience	240	0.5
Clean	3104	5.9	Airport	620	1.2	Internet	344	0.7	Suite	236	0.4
Staff	2898	5.5	Desk	609	1.2	Trip	328	0.6	Money	233	0.4
Location	2865	5.4	View	569	1.1	Pay	320	0.6	Carpet	233	0.4
Comfortable	2168	4.1	Recommend	532	1.0	Door	317	0.6	Courteous	233	0.4
Service	1707	3.2	Noise	493	0.9	Shops	316	0.6	City	231	0.4
Friendly	1614	3.1	Quiet	486	0.9	Sleep	303	0.6	Expensive	223	0.4
Close	1594	3.0	Food	468	0.9	Business	301	0.6	Dirty	221	0.4
Breakfast	1524	2.9	Distance	464	0.9	Complaint	299	0.6	Renovated	219	0.4
Helpful	1378	2.6	Shuttle	447	0.8	Shower	296	0.6	Tub	217	0.4
Bed	1334	2.5	Street	429	0.8	Family	294	0.6	Safe	216	0.4
Price	1321	2.5	Shopping	419	0.8	Value	290	0.5	Far	214	0.4
Restaurants	1153	2.2	Maintained	417	0.8	Cheap	288	0.5	Air	213	0.4
Walking	1011	1.9	Beach	398	0.8	Smelled	284	0.5	Refrigerator	205	0.4
Area	863	1.6	Access	398	0.8	Kids	258	0.5	Quality	203	0.4
Parking	802	1.5	Park	385	0.7	Tv	256	0.5	Decor	201	0.4
Bathroom	764	1.4	Floor	373	0.7	Attractions	248	0.5	Wait	200	0.4
Pool	716	1.4	Check in	369	0.7	Water	247	0.5	Freeway	198	0.4
Free	712	1.3	Spacious	365	0.7	Coffee	244	0.5	Elevator	196	0.4
Convenient	708	1.3	Bar	358	0.7	Amenities	244	0.5	Accommodation	114	0.2

"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 2, six meaningful factors consisting of 34 words out of the final 80 words in Table 1 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 rep-resent the common semantic spaces in customer reviews. Since the loadings were relatively low (compared to factor analysis con-ducted using established metric scales), the cutoff loading was set at (±) 0.30 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

Table 2. 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]	0.436					
Smelled (0.5)	0.423					
Dirty (0.4)	0.395					
Price (2.5)	0.369					
Cheap (0.5)	0.354					
Carpet (0.4)	0.349					
Sleep (0.6)	0.323					
Expensive (0.4)	−0.313					
Shopping (0.8)	−0.326					
View (1.1)	−0.377					
Restaurants (2.2)	−0.387					
Distance (0.9)	−0.459					
Location (5.4)	−0.492					
Walking (1.9)	−0.496					
Deals						
Breakfast (2.9)		0.517				
Airport (1.2)		0.433				
Free (1.3)		0.435				
Comfortable (4.1)		0.409				
Shuttle (0.8)		0.393				
Amenities						
Close (3.0)			0.390			
Beach (0.8)			−0.366			
Pool (1.4)			−0.533			
Family friendliness						
Family (0.6)				0.509		
Kids (0.5)				0.483		
Attractions (0.5)				0.338		
Suite (0.4)				0.313		
Service (3.2)				−0.338		
Core product						
Room (10.7)					0.552	
Bathroom (1.4)					0.420	
Bed (2.5)					0.322	
Spacious (0.7)					0.302	

Continued

Staff						
Helpful (2.6)						−0.462
Friendly (3.1)						−0.511
Staff (50.5)						−0.517
Eigenvalue	4.55	3.65	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 1).

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 "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 3 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 = 0.01 level, with the first two factors, Hybrid and Deals having the largest standardized coefficients of −0.567 and 0.506, 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 positive, it means that a high satisfaction rating is associated with the mentions of words about free services (*i.e.*, breakfast and air-port 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"

Table 3. Results of linear regression analysis.

Model	Unstandardized coefficients		Standardized coefficients	*t*	Sig.
	B	Std. error	Beta		
(Constant)	4.023	0.013		298.410	0.000
Hybrid	−0.293	0.013	−0.576	−21.714	0.000
Deals	0.258	0.013	0.506	19.086	0.000
Amenities	−0.015	0.013	−0.029	−1.076	0.282
Family friendliness	0.076	0.013	0.149	5.606	0.000
Core Product	0.063	0.013	0.123	4.641	0.000
Staff	0.044	0.013	0.086	3.242	0.001

Dependent variable: average customer rating; Adjusted R square: 0.629.

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 and Implications

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. 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 [2]. 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 [13], 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 [27]. 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 [28] and its variants in the field of hospitality and tourism [29]. 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 [20].

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 [27]. These insights also attest to the capa-

bilities 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 [30], when hotel guest satisfaction is being examined as the dependent variable, an R square value between 0.50 and 0.60 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 [22], 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.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Browning, V., So, K.K.F. and Sparks, B. (2013) The Influence of Online Reviews on Consumers' Attributions of Service Quality and Control for Service Standards in Hotels. *Journal of Travel & Tourism Marketing*, **30**, 23-40. https://doi.org/10.1080/10548408.2013.750971

[2] George, G., Haas, M.R. and Pentland, A. (2014) Big Data and Management. *Academy of Management Journal*, **57**, 321-326. https://doi.org/10.5465/amj.2014.4002

[3] Ye, Q., Law, R. and Gu, B. (2009) The Impact of Online User Reviews on Hotel Room Sales. *International Journal of Hospitality Management*, **28**, 180-182. https://doi.org/10.1016/j.ijhm.2008.06.011

[4] Wu, C.H.J. and Liang, R.D. (2009) Effect of Experiential Value on Customer Satisfaction with Service Encounters in Luxury-Hotel Restaurants. *International Journal of Hospitality Management*, **28**, 586-593. https://doi.org/10.1016/j.ijhm.2009.03.008

[5] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, **457**, 1012-1014. https://doi.org/10.1038/nature07634

[6] Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data: Provocations for

a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, **15**, 662-679. https://doi.org/10.1080/1369118X.2012.678878

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

[8] Schumaker, R.P. and Chen, H. (2009) Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZF in Text System. *ACM Transactions on Information Systems*, **27**, 1-19. https://doi.org/10.1145/1462198.1462204

[9] Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008) More than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, **63**, 1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

[10] Ghose, A. and Ipeirotis, P.G. (2011) Estimating the Helpfulness and Economic Impact of product Reviews: Mining Text and Reviewer Characteristics. Knowledge and Data Engineering. *IEEE Transactions on Knowledge and Data Engineering*, **23**, 1498-1512. https://doi.org/10.1109/TKDE.2010.188

[11] Abrahams, A.S., Jiao, J., Wang, G.A. and Fan, W. (2012) Vehicle Defect Discovery from Social Media. *Decision Support Systems*, **54**, 87-97. https://doi.org/10.1016/j.dss.2012.04.005

[12] Ghose, A., Ipeirotis, P.G. and Li, B. (2012) Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd Sourced Content. *Marketing Science*, **31**, 493-520. https://doi.org/10.1287/mksc.1110.0700

[13] Pang, B. and Lee, L. (2008) Opinion Mining and Sentiment Analysis. *Found. Trends Inform. Retr.*, **2**, 1-135. https://doi.org/10.1561/1500000011

[14] Schmunk, S., Höpken, W., Fuchs, M. and Lexhagen, M. (2013) Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC. In: Xiang, Z. and Tussyadiah, I., Eds., Information and Communication Technologies in Tourism 2014. Springer International Publishing, New York, 253-265. https://doi.org/10.1007/978-3-319-03973-2_19

[15] Hunt, J.D. (1975) Image as a Factor in Tourism Development. *Journal of Travel Research*, **13**, 3-7. https://doi.org/10.1177/004728757501300301

[16] Oliver, R.L. (1981) Measurement and Evaluation of Satisfaction Processes in Retail Settings. *Journal of Retailing*, **57**, 25-48.

[17] Engel, J.F., Blackwell, R.D. and Miniard, P.W. (1990) Consumer Behavior. 6th Editon, Dryden Press, Hinsdale.

[18] Yoon, Y. and Uysal, M. (2005) An Examination of the Effects of Motivation and Satisfaction on Destination Loyalty: A Structural Model. *Tour Manager*, **26**, 45-56. https://doi.org/10.1016/j.tourman.2003.08.016

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

[20] Chathoth, P., Altinay, L., Harrington, R.J., Okumus, F. and Chan, E.S. (2013) Co-Production versus Co-Creation: A Process Based Continuum in the Hotel Service Context. *International Journal of Hospitality Management*, **32**, 11-20. https://doi.org/10.1016/j.ijhm.2012.03.009

[21] Lockyer, T. (2005) The Perceived Importance of Price as One Hotel Selection Dimension. *Tour Manager*, **26**, 529-537. https://doi.org/10.1016/j.tourman.2004.03.009

[22] Crotts, J.C., Mason, P.R. and 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. *Journal of Travel Research*, **48**, 139-151. https://doi.org/10.1177/0047287508328795

[23] Oh, H. (1999) Service Quality, Customer Satisfaction, and Customer Value: A Holistic Perspective. *International Journal of Hospitality Management*, **18**, 67-82. https://doi.org/10.1016/S0278-4319(98)00047-4

[24] Pan, B., MacLaurin, T. and Crotts, J.C. (2007) Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research*, **46**, 35-45. https://doi.org/10.1177/0047287507302378

[25] Serra Cantallops, A. and Salvi, F. (2014) New Consumer Behavior: A Review of Research on eWOM and Hotels. *International Journal of Hospitality Management*, **36**, 41-51. https://doi.org/10.1016/j.ijhm.2013.08.007

[26] Gretzel, U. and Yoo, K.H. (2008) Use and Impact of Online Travel Reviews. In: *Information and Communication Technologies in Tourism* 2008, Springer, Vienna, 35-46. https://doi.org/10.1007/978-3-211-77280-5_4

[27] Stringam, B.B. and Gerdes Jr., J. (2010) An Analysis of Word-of-Mouse Ratings and Guest Comments of Online Hotel Distribution Sites. *Journal of Hospitality Marketing & Management*, **19**, 773-796.

[28] Herzberg, F. (1966) Work and the Nature of Man. World Publishing, Cleveland.

[29] Noe, F.P. and Uysal, M. (1997) Evaluation of Outdoor Recreational Settings: A Problem of Measuring User Satisfaction. *Journal of Retailing and Consumer Services*, **4**, 223-230. https://doi.org/10.1016/S0969-6989(96)00030-6

[30] Lewis, R.C. (1985) Getting the Most from Marketing Research: Part V. Predicting Hotel Choice: The Factors Underlying Perception. *The Cornell Hotel and Restaurant Administration Quarterly*, **25**, 82-96. https://doi.org/10.1177/001088048502500415