Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation

Yue Guo [a], Stuart J. Barnes [b, *], Qiong Jia [a]

[a] Hohai University, Nanjing City, Jiang Su Province 21110, China
[b] School of Management and Business, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, United Kingdom

HIGHLIGHTS

- Extraction of dimensions of visitor satisfaction from online hotel reviews.
- Uses latent dirichlet analysis on big data to identify the voice-of-the customer.
- Identification of 19 controllable dimensions for hotel-customer interaction.
- Heterogeneity found amongst different visitor demographic segments.
- Perceptual mapping identifies key dimensions according to hotel star-rating.

ARTICLE INFO

ABSTRACT

Consumer-generated content has provided an important new information medium for tourists, throughout the purchasing lifecycle, transforming the way that visitors evaluate, select and share experiences about tourism. Research in this area has largely focused on quantitative ratings provided on websites. However, advanced techniques for linguistic analysis provide the opportunity to extract meaning from the valuable comments provided by visitors. In this paper, we identify the key dimensions of customer service voiced by hotel visitors use a data mining approach, latent dirichlet analysis (LDA). The big data set includes 266,544 online reviews for 25,670 hotels located in 16 countries. LDA uncovers 19 controllable dimensions that are key for hotels to manage their interactions with visitors. We also find differences according to demographic segments. Perceptual mapping further identifies the most important dimensions according to the star-rating of hotels. We conclude with the implications of our study for future research and practice.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Prior studies in marketing and consumer behavior have defined customer satisfaction as a customer's subjective evaluation of a service or product provided based on expectations and actual performance (Anderson, Fornell, & Lehman, 1994; Oliver, 1980; Woodruff, Cadotte, & Jenkins, 1983). Customers evaluate the degree of satisfaction based on their perceptions of the attributes of hotels that they deem most important. In other words, these attributes of hotels represent dimensions of satisfaction. Numerous studies have proposed that customer satisfaction plays an important role in motivating customers' behavioral loyalty, such as giving positive reviews, returning, or making a recommendation (e.g., Hallowell, 1996; Hui, Wan, & Ho, 2007; Kim, Ng, & Kim, 2009a). Multiple factors contribute to the formation of consumer satisfaction (e.g., price, service quality, and product quality), and it is thus a multidimensional construct consisting of different aspects or subconstructs, in a similar manner to service quality (Klein & Leffler, 1981; Mitra & Golder, 2006; Tellis & Johnson, 2007). Prior studies typically rely on traditional qualitative, quantitative or mixed methods (e.g., questionnaire survey and focus groups) to identify the dimensions of satisfaction, and subsequently develop empirical measurement scales. These traditional research methods require researchers to seek an effective trade-off between the cost of sample collection and estimation performance. Existing studies on

* Corresponding author.
E-mail addresses: yueggcn@aliyun.com (Y. Guo), stuart.barnes@kcl.ac.uk (S.J. Barnes), jiaqionghit@163.com (Q. Jia).

satisfaction are empirically examined based on limited samples during a specific period. Moreover, initial measurement items and survey questions tend to be developed based on the knowledge of researchers on related industries (e.g., the hospitality industry). At a result, inconsistent measurement items and questions are often created and used in prior studies (Barsky, 1992; Danaher & Haddrell, 1996; Fornell, 1992).

The Internet has fostered a rapid rise in user-generated content (UGC), particularly alongside the widespread diffusion of Web 2.0 technologies. Tourists can now share their experiences and give specific suggestions to others for hotels, restaurants and attractions (e.g., via comments on customer service, car parking and cleanliness) (Sotiriadis & van Zyl, 2013, Sparks & Browning, 2011; Vermeulen & Seegers, 2009; Ye, Law, & Gu, 2009; Ye, Zhang, & Law, 2009). Extant literature has shown that online customer reviews can be used as a major information source for researchers and practitioners that can help in correctly understanding consumer preferences and demand: for example, to predict financial performance or attempt to increase sales (e.g., Chau & Xu, 2012; Chevalier & Mayzlin, 2006; Clemons, Gao, & Hitt, 2006; Ghose & Ipeirotis, 2011; Liu, 2006; Ye, Law, Li, & Li, 2011). Online customer reviews can empower individuals to bypass unclear and inaccurate product or service descriptions and rely directly on the first-hand usage experiences of other consumers, particularly in the case of high involvement products (e.g., vehicles). Moreover, some firms actively encourage their customers to submit online opinions about their products or services, e.g., by offering vouchers or discounts.

UGC may be considered as spontaneous, insightful and passionate feedback provided by consumers that is widely available, free or low cost, and easily accessible anywhere, anytime. Large volumes of data, as represented in the continuous stream of UGC over time, provides practical input (i.e., know-how, know-what) to augment traditional research methods for identifying important issues. Latent dimensions, such as social status, are variables that consumers may not explicitly mention, but that capture or represent a large number of attributes, often indirectly from other indicators (e.g., income and occupation). As a consequence, in the past decade, there have been an increasing number of studies examining the phenomenon of online consumer reviews (e.g., Clemons et al., 2006; Dellarocas, Zhang, & Awad, 2007; Ho-Dac, Carson, & Moore, 2013). Specifically, regarding the hospitality and tourism field, we propose that UGC provides a rich source of data to extract the dimensions of customer satisfaction. Hundreds of thousands of community members may contribute to creating online content, thereby creating the "wisdom of crowds" (Surowiecki, 2005). Thus, UGC can serve as a useful source of information for enterprises that care about consumers' demands, particularly in the hospitality industry (e.g. hotels and restaurants). Appendix A provides a brief summary of the recent empirical literature on UGC from multiple academic disciplines, including travel and tourism, marketing, and information systems, since research on UGC in tourism is a recent phenomenon (e.g., Li, Ye, & Law, 2013; Liu & Park, 2015; Park & Nicolau, 2015). Our review includes influential studies and is intended to be representative rather than exhaustive. Despite the growing significance of online reviews, and in the face of concerns voiced about them, we argue that the extant literature mainly focuses upon examining the impact of online ratings; thus, the literature largely ignores online reviews, which we consider potentially more valuable to academics and practitioners. For example, most prior tourism studies have principally focused on the impact of online ratings on hotel sales (e.g., Xie, Miao, Kuo, & Lee, 2011; Ye , Law, & Gu, 2009; Ye, Law, Yi, & Li, 2011). Compared to online ratings, which are numerical and easily understood, online reviews are text-based and often comprise of large information

repositories beyond the analytical capabilities of traditional econometric and statistical methods. Thus far, there is limited empirical evidence from large-scale online reviews to help in understanding consumer satisfaction and its antecedents.

An emerging stream of research in the hospitality industry has attempted to understand customer satisfaction from the content of online reviews. For example, Li et al. (2013) exploit 42,886 online reviews of 774 star-rated hotels in Beijing and use content analysis to identity the determinants of satisfaction. Lu and Stepchenkova (2012) analyzed 373 reviews extracted from TripAdvisor based on content analysis to identity satisfaction attributes. Levy et al. (2013) carry out a content analysis of complaints within one-star online hotel reviews from ten popular review websites to understand customer satisfaction. They content analyzed 225 managerial responses to these one-star reviews. Relative to these prior studies, the proposed framework in the current study differs in three major ways. First, in this research, we attempt investigate online reviews and online ratings together. On the one hand, more effort has to be put into the analysis of large-scale online review contents in order to explore fully the potential of the data to identify the antecedents of satisfaction. The study uses topic modelling: advanced software and mathematical techniques developed in the fields of natural language processing and data mining. On the other hand, a stepwise regression analysis of a large volume of numerical ratings is carried out to verify the validity of important dimensions proposed by prior empirical studies on satisfaction of hotel customers. These studies have principally used questionnaire surveys or experiments to understand the antecedents of satisfaction (e.g., Min, Lim, & Magnini, 2015; Wu & Liang, 2009). Such a combined analysis of online ratings and online reviews enables us to further link specific dimensions extracted from large textual contents with abstract-level factors identified by prior studies, such as service quality. Second, LDA uses an unsupervised Bayesian learning algorithm to capture effectively context-specific dimensions and does not make any assumption about the distribution of online reviews or grammatical attributes of language. Consequently, relative to prior methods for text analysis, LDA can complete many steps of the textual analysis with little human intervention, even labeling dimensions, and is more suitable for dealing with large and unstructured online reviews, thus creating meanings that are more realistic. Finally, this study demonstrates the method on a relatively broad sample of more than 200,000 online reviews of hotels located in more than 100 cities in 16 countries, which enables us to make more reliable generalizations than prior studies.

In sum, an important contribution of this research is that we empirically develop and identify the dimensions of satisfaction based on big data from UGC including numerical and textual information. Thus, the dimensions from the data provide a genuine "voice of the customer" (Griffin & Hauser, 1993), generally understood as:

> "a complete set of customer wants and needs; expressed in the customer's own language; organized the way the customer thinks about, uses and interacts with the product and service; and prioritized by the customer in terms of both importance and performance — in other words current satisfaction with existing alternatives" (Katz, 2011, p. 34).

In this study, we attempt to mine the sensitive and important factors influencing consumer satisfaction through UGC. By extracting value from UGC, we believe we will be able to hear the voice of the customer more correctly and effectively, providing practical help for business owners and investors.

Specifically, the goal of this study is to answer the following questions: