A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism

Zheng Xiang [a, *], Qianzhou Du [b], Yufeng Ma [c], Weiguo Fan [d]

[a] Department of Hospitality and Tourism Management, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA
[b] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA
[c] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
[d] Center for Business Intelligence & Analytics, Department of Accounting and Information Systems, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA

HIGHLIGHTS

- We applied text analytics to compare three major online review platforms, namely, TripAdvisor, Expedia, and Yelp.
- Findings show discrepancies in the representation of hotel product on these platforms.
- Information quality, measured by linguistic and semantic features, sentiment, rating, and usefulness, varies considerably.
- This study is the first to comparatively explore data quality in social media studies in hospitality and tourism.
- This study highlights methodological challenges and contributes to the theoretical development of social media analytics.

ARTICLE INFO

ABSTRACT

Online consumer reviews have been studied for various research problems in hospitality and tourism. However, existing studies using review data tend to rely on a single data source and data quality is largely anecdotal. This greatly limits the generalizability and contribution of social media analytics research. Through text analytics this study comparatively examines three major online review platforms, namely TripAdvisor, Expedia, and Yelp, in terms of information quality related to online reviews about the entire hotel population in Manhattan, New York City. The findings show that there are huge discrepancies in the representation of the hotel industry on these platforms. Particularly, online reviews vary considerably in terms of their linguistic characteristics, semantic features, sentiment, rating, usefulness as well as the relationships between these features. This study offers a basis for understanding the methodological challenges and identifies several research directions for social media analytics in hospitality and tourism.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

There is a growing literature on social media analytics that combines Web crawling, computational linguistics, machine learning and statistical techniques to collect, analyze, and interpret the so-called big data for business purposes such as tracking trending topics and popular sentiments as well as identifying opinions and beliefs about products (Fan & Gordon, 2014; Lazer et al., 2009). Particularly, online consumer reviews, widely considered a rich data source that reflects consumer experiences and evaluation of products, have been studied to understand a range of research problems in hospitality and tourism (Schuckert, Liu, & Law, 2015b). Studies using online reviews usually employ a sample of review (and related) data, large or small, to extract features or measures that allow the researcher to detect, describe or predict patterns that are meaningful from theoretical or practical perspectives. This literature complements conventional approaches that primarily rely on surveys, personal interviews and other communication-based methods and represents a promising research direction by taking advantage of the abundant, readily available data resources (Xiang, Schwartz, Gerdes, & Uysal, 2015). While this line of research has generated novel insights into hospitality and tourism management, existing studies are limited in that 1) they tend to use a single data source for online reviews and 2) the quality of data is

* Corresponding author.
E-mail addresses: philxz@vt.edu (Z. Xiang), qiand12@vt.edu (Q. Du), yufengma@vt.edu (Y. Ma), wfan@vt.edu (W. Fan).

largely anecdotal and often based upon the popularity of the websites from which the data were collected, which substantially limits their generalizability and contribution to knowledge.

With this in mind, this study comparatively examines three major online review platforms, namely TripAdvisor, Expedia, and Yelp, in terms of information quality related to online reviews in these websites with the goal to provide a basis for understanding the methodological challenges and for identifying opportunities for the development of social media analytics in hospitality and tourism. The rest of the paper is organized as follows: the next section, Research Background, reviews related literature to provide the motivations for the present study. In Research Design and Analytical Framework, we outline our methodological approaches and describe key measures and methods used to assess the three platforms with specific research questions. In Data Collection and Analysis, we describe the data collection process and explain, in details, the text analytics procedures to develop key metrics to describe review characteristics as well as statistical analyses conducted to compare and contrast the three platforms. Then, research findings are presented followed by a discussion on implications for both research and practice. Finally, conclusions are drawn, and limitations and directions for future research are discussed.

2. Research background

Big data analytics has been touted as a new research paradigm that utilizes diverse sources of data and analytical tools to make inferences and predictions about reality (Boyd & Crawford, 2012; Mayer-Schönberger & Cukier, 2013). Particularly, with increasingly powerful natural language processing and machine learning capabilities, textual contents from the Web provide a huge shared cognitive and cultural context and, thus, have been analyzed in many application domains (Halevy, Norvig, & Pereira, 2009). However, in recent years there have been growing criticism and concerns about the data-driven approach especially those using online user-generated contents as research data. For example, Ekbia et al. (2015) discuss some of the epistemological dilemmas in existing big data analytics, including the validity of claims about causal relationships, as opposed to mere statistical correlations, within the data. Others (e.g., Frické, 2015) challenge the nature of inductive reasoning in big data analytics and suggest that there are potential hazards in making generalizable claims. Particularly, Ruths and Pfeffer (2014) argue that studies using social media data should be aware of a number of validity problems such as platform biases (e.g., platform design, user base, and platform specific behavior), data availability biases, and data authenticity issues. Tufekci (2014) specifically highlights the conceptual and methodological challenges in social media studies, particularly sampling biases arising from using a single platform as data source due to the sociocultural complexity of user behavior and unrepresentative sampling frame, which may complicate the interpretation of research findings. Importantly, she argues that social media platforms are comparable to specimens in biological research wherein they are selected for certain characteristics suitable for laboratory examinations at the expense of illuminating other potentially important features. As such, Frické (2015) suggests that correlations found in many of existing big data studies might only be considered "candidate" solutions to the problems at hand. Ruths and Pfeffer (2014), among others, call for the use of a variety of triangulation approaches for big data analytics such as applying the same methods to examine the performance on two or more distinct data sets when studying a new social phenomenon.

The online eco-system in hospitality and tourism is vast, complex, and diverse; so are online review platforms, which range from community-based sites such as LonelyPlanet, Tripadvisor and Yelp

to transaction-based online travel agencies (aka OTAs) like Expedia and Bookings.com where reviews are incorporated as electronic word-of-mouth (Gligorijevic, 2016). Although they may all be considered part of social media with the common goal to assist consumer decision making by providing trusted, shared social knowledge, these platforms are complex sociocultural and economic systems that reflect different business models, different technological affordances, and different user segments/bases and power distribution in the online eco-system (Jeacle & Carter, 2011; Scott & Orlikowski, 2012). For example, TripAdvisor, by incorporating a variety of user data and information tools, represents various actors, resources and, importantly, business models through its website (Yoo, Sigala, & Gretzel, 2016). Recent market dynamics such as Expedia's takeover of Travelocity and Orbitz have created a new power structure within the online eco-system with the emergence of a potentially dominant social knowledge base (see http://time.com/money/3707551/expedia-orbitz-impact-travelers). From the business viewpoint, online reviews including their peripheral cues, such as user-supplied photos and the reviewer's personal information, are intended as means of persuasive communication in order to build credibility and influence user behavior (Sparks, Perkins, & Buckley, 2013; Zhang, Zhang, & Yang, 2016). Therefore, the selection, ranking, and display of online reviews reflect the platform's strategy to maximize these effects on its targeted audience. Also, the contribution of online reviews is a self-selection process, which contributes to quality differentiation over the life cycle of product reviews (Li & Hitt, 2008). Mkono and Tribe (2016), in a recent study, show that users on travel-related social media sites not only are product evaluators but also may play additional, important roles such as online activist, troll, social critic, information seeker and socialite. Furthermore, these issues are confounded with the long-standing concerns about the authenticity of online reviews (Luca & Zervas, 2015). Therefore, social media research using online review data must be cognizant of the nuances in these data sources in order to make conscious, appropriate methodological decisions when considering the representativeness and quality of the data.

Hospitality and tourism appears to be an ideal application field of social media analytics with tremendous growth and potential. For example, a recent study (Lu & Stepchenkova, 2015) cited over 100 papers primarily focusing on user-generated contents published in hospitality and tourism journals in the previous 10 years. Schuckert et al. (2015b) cited 50 articles related to online reviews published within and outside the field of hospitality and tourism, indicating the growing interest in understanding the impact of online reviews. For the present study, we examined the literature that specifically used online reviews as data in hospitality and tourism. Table 1 lists a sample of recent publications in six leading tourism and hospitality journals, namely *Tourism Management*, *Annals of Tourism Research*, *Journal of Travel Research*, *International Journal of Hospitality Management*, *Cornell Hospitality Quarterly*, and *Journal of Tourism and Hospitality Research*. While there is a considerable amount of publications elsewhere within and outside hospitality and tourism and, therefore, this compilation by no means represents a full picture of the literature, these journals were selected as the sampling frame because of their high influence in the field (McKercher, Law, & Lam, 2006). As shown in the table, these studies all collected and analyzed online reviews (and associated information) to address a variety of research problems such as travel motivation (e.g., Pearce & Wu, 2015), opinions and sentiments related to hospitality products (e.g., Crotts, Mason, & Davis, 2009; Levy, Duan, & Boo, 2013; Xiang et al., 2015), impact of online reviews on hotel business performance (e.g., Melián-González, Bulchand-Gidumal, & López-Valcárcel, 2013; Xie, Zhang, & Zhang, 2014; Ye, Law, & Gu, 2009), and the nature and utilities of online reviews as data (e.g., Fang, Ye,