Contents lists available at ScienceDirect

Tourism Management Perspectives

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

Studying the association of online brand importance with museum visitors: An application of the semantic brand score

Andrea Fronzetti Colladon[a,*], Francesca Grippa[b], Rosy Innarella[c]

[a] University of Perugia, Department of Engineering, Via G. Duranti 93, 06125 Perugia, Italy
[b] Northeastern University, College of Professional Studies, Boston, 360 Huntington Avenue, MA 02115 Boston, MA, United States of America
[c] University of Rome Tor Vergata, Department of Enterprise Engineering, Viale del Politecnico 1, 00133 Rome, Italy

ARTICLE INFO

Keywords:
Big data
Social media
Semantic brand score
Brand importance
Museum marketing

ABSTRACT

This paper explores the association between brand importance and growth in museum visitors. We analyzed 10 years of online forum discussions and applied the Semantic Brand Score (SBS) to assess the brand importance of five European Museums. Our Naive Bayes and regression models indicate that variations in the combined dimensions of the SBS (prevalence, diversity and connectivity) are aligned with changes in museum visitors. Results suggest that, in order to attract more visitors, museum brand managers should focus on increasing the volume of online posting and the richness of information generated by users around the brand, rather than controlling for the posts' overall positivity or negativity.

1. Introduction

Competition in the tourism industry is increasingly based on high-volume data that is immediately available for travelers and tourism industry operators. In this increasingly knowledge intensive industry, big data represents an asset for various stakeholders, reducing information asymmetry for customers and increasing flexibility and responsiveness for organizations. Big data requires specific technology and analytical methods for its transformation into value (De Mauro et al., 2015). While recent studies used Google Trends Data to better understand tourist interests and intentions (Li, Pan, Law, & Huang, 2017; Padhi & Pati, 2017; Park, Lee, & Song, 2017), or focused on the analysis of online reviews (Fang, Ye, Kucukusta, & Law, 2016; Lee, Law, & Murphy, 2011; Wong & Qi, 2017), a less explored area is the analysis of the content extracted from online forums with the goal to predict museum visitors. Analyzing the content exchanged by users on sites such as TripAdvisor can help design promotional campaigns and brand awareness strategies that could inform and guide users' purchasing behaviors (Banerjee & Chua, 2016).

Leveraging the power of big data has the potential to reveal patterns and trends that are beneficial to several stakeholders in the tourism industry (Mandal, 2018; Xiang, 2018). Individual travelers can make decisions faster and use more complete and diversified information, which impacts the quality of their experience when choosing a destination to visit, comparing prices and building expectations for an upcoming trip (Leung, Law, van Hoof, & Buhalis, 2013).

Social media and online review sites support information search, decision-making and knowledge exchange for tourists and represent an opportunity for companies in the tourism industry to learn more about needs and find new ways to meet travelers' expectations (Gavilan, Avello, & Martinez-Navarro, 2018; Moro, Rita, & Coelho, 2017). Online travel forums are the ideal space for tourists to find answers to specific questions and link to resources to help them make the right decisions as they plan their travels (Hwang, Jani, & Jeong, 2013). Reviewers often share advice on practical matters, motivated by a desire for community empowerment, social support and joint-affirmation (Munar & Ooi, 2012).

The benefits of leveraging big data analytics to support strategic decision making in tourism destination management have emerged only over the past few years (Miah, Vu, Gammack, & McGrath, 2017), representing an interesting gap that this research would like to address. We explore the potential application of a big data method to extract information from online travel forums. The goal is to evaluate the association between museum visitors and museum brand importance, by testing the potential value of new indicators that could be used to improve existing forecasting models.

We applied a measure defined Semantic Brand Score (SBS) that has been used to assess brand importance in other industries. SBS combines methods of semantic analysis and social network analysis to study large text corpora, across products, markets and languages (Fronzetti

[*] Corresponding author.
E-mail addresses: andrea.fronzetticolladon@unipg.it (A. Fronzetti Colladon), f.grippa@northeastern.edu (F. Grippa), rosy.innarella@uniroma2.it (R. Innarella).

https://doi.org/10.1016/j.tmp.2019.100588
Received 14 January 2019; Received in revised form 12 September 2019; Accepted 18 October 2019

Colladon, 2018). Aligned with the work of Fronzetti Colladon (2018), we conceptualize online brand importance via three dimensions: prevalence, diversity and connectivity. These dimensions reflect respectively the frequency of use of a brand name (prevalence), such as a museum name, the heterogeneity of its textual brand associations (diversity) and its embeddedness at the core of a discourse (connectivity). To assess the benefits of adopting such a method to measure brand importance and relate it to museum visitors, in this paper we focus on five popular museums located in Italy, France and Hungary.

Consistently with this conceptualization of brand importance, our study aims to address the following research question: is museum brand importance associated with variations of visitors over time?

The paper is organized as follows. Next section offers an overview of the theoretical background on social media, big data and text mining in the tourism industry, providing a conceptual framework that highlights the hypotheses. The third section describes the research design and methodology, also describing sample and data collection strategy. After illustrating our findings, the final section is devoted to discussing results and their implications.

2. Theoretical background: social media, brands and tourism

Over the past ten years we have seen an increasing number of studies focused on the effect of social media on tourist decisions (Jacobsen & Munar, 2012; Miguéns, Baggio, & Costa, 2008), suggesting a positive linkage between perceived quality, electronic word of mouth, brand image, and brand performance (Barnes, Mattsson, & Sørensen, 2014). The high-context interactions offered by social media platforms impact consumers' attitudes and behaviors, leading them to perceive the brand in more positive terms, which ultimately increases purchase intentions (Kim & Ko, 2012). By interacting with others on a forum, users can reduce misunderstanding on what a brand offers, they can change their attitude from negative or neutral to positive, as well as receive additional details that will lead them to finalize the purchase. Recent studies focused on specific aspects of brand equity, such as brand popularity (Gloor, 2017; Gloor, Krauss, Nann, Fischbach, & Schoder, 2009). Others applied methodologies to web data and focused on social interaction via social media (De Vries, Gensler, & Leeflang, 2012), or studied links among webpages or user generated content (Yun & Gloor, 2015).

Brand-based social media activities generate online or electronic word of mouth that could improve the understanding of products and services thanks to the exchange of ideas among users, which improves marketing productivity and performance (Filieri & McLeay, 2014; Keller, 1993; Torres, Singh, & Robertson-Ring, 2015). A recent study on the impact of user interactions in social media (Hutter, Hautz, Dennhardt, & Füller, 2013) found that users' engagement with a Facebook page positively impact their brand awareness, word-of-mouth engagement and even purchase intentions. In a study focused on user-generated content and the effects of electronic word-of-mouth on hotel online bookings, Ye, Law, Gu, and Chen (2011) found that online reviews have a significant impact on online sales, with a 10% increase in traveler review ratings boosting online bookings by more than 5%. Yang, Pan, Evans, and Lv (2015) used web search query volume on Google and Baidu to predict visitor numbers for a popular tourist destination in China and demonstrated the predictive power of using search engines in understanding the travel process of tourists. Other studies explored the impact of good vs. bad ratings during the first stage of the decision-making process when travelers book a hotel, and found that web users tend to select hotels that have better ratings (Gavilan et al., 2018; Sparks & Browning, 2011). Similarly, Neirotti, Raguseo, and Paolucci (2016) demonstrated how recommendations posted on social media by peers can positively influence travelers in choosing hotels and destinations that are consistent with their preferences and attitudes. These approaches, however, have some limitations when applied to contexts containing text that is not associated to a social interaction (for example a press release). In addition, we found a limited number of studies that linked semantic analysis with factors impacting brand equity.

2.1. Text mining in tourism

Assessing brand equity and importance has been usually done via market surveys administered to consumers and other stakeholders, or via financial methods (Belén del Río, Vázquez, & Iglesias, 2001; Lassar, Mittal, & Sharma, 1995). Surveys and traditional financial methods have limitations due to perception biases, sampling methods and excessive dependence on historical variables. An increasing number of studies are adopting text mining techniques and sentiment analysis approaches to study the informative contribution of travelers and users in online forums, with only few of them focused on museum visitors (Volcheck, Song, Law, & Buhalis, 2018). Aggarwal, Vaidyanathan, and Venkatesh (2009) extracted information using the Google search engine and lexical text analysis to explore online brand representations, by examining the association between brands and a selection of adjectives or descriptors. Others applied text analytics to online customer reviews collected from Expedia.com to understand hotel guest experience and its relationships with guest satisfaction (Xiang, Schwartz, Gerdes, & Uysal, 2015). Their results show that the association between satisfaction rating and guest experience is strong, and that a general pattern can be observed between customers' use of particular words to describe the experience and the quality of service provided. Another interesting example of text mining applied to understand tourism demand is offered by the application of content analysis to 220 samples of Lonely Planet postings to assess the messages' functional information (Hwang et al., 2013).

Researchers have analyzed users' interactions by mining forum posts, mailing list archives, hyperlink structure of homepages, or co-occurrence of text in documents, though fewer have explored how content analysis on social media could be used in an integrated way to understand brand importance online. Text mining of online tourism reviews offers invaluable - and otherwise difficult to collect - review evaluations supporting comparative analysis (Bucur, 2015; Hu & Liu, 2004).

All the empirical studies on travelers' online behavior and its impact on economic performance that we have presented thus far have been heterogeneous and focused on a multiplicity of big data approaches. Most of the text mining systems and approaches developed in the past few years are based on an extraction of reviews from page content, and then use algorithms or text mining modules to process the content through a classification of reviews as positive, negative and neutral (Capodieci, Elia, Grippa, & Mainetti, 2019; Zhang, Fuehres, & Gloor, 2011). The framework that we use in this study (Semantic Brand Score) goes beyond the textual classification of words or comments on social media, and incorporates new metrics of text analysis with indicators developed in the fields of social network analysis and semantic analysis (Fronzetti Colladon, 2018).

2.2. An integrated framework to study brand importance

The Semantic Brand Score (SBS) is a comprehensive framework based on widely accepted brand equity models (Keller, 1993; Wood, 2000) that evaluates brand importance using a composite approach that goes beyond counting the number of likes to Facebook brand pages or the number and valence of comments on social media. We will use the SBS composite indicator as the basis for our conceptual framework. The SBS is calculated based on three dimensions: prevalence, diversity and connectivity. Partially connected to the dimension of brand awareness (Aaker, 1996), prevalence represents the frequency with which the brand name appears in a set of text documents: the more frequently a brand is mentioned, the higher its prevalence. Prevalence looks at the frequency by which a museum name is mentioned in a discourse, and can be intended as a proxy for brand awareness. Keller's definition of brand equity and brand awareness includes the concept of differential response to knowledge of a brand name, suggesting that brand awareness is the starting point to building a positive image (Keller, 1993). The second SBS dimension, diversity, is related to the concepts of lexical diversity (McCarthy & Jarvis, 2010) and word co-occurrences (Evert, 2005). It measures the heterogeneity of the words co-occurring with a brand, assigning higher diversity to brands embedded in a richer discourse. '*A brand could be mentioned frequently in a discourse, thus having a high prevalence, but always used in conjunction with the same words, being limited to a very specific context*' (Fronzetti Colladon, 2018, p. 152). The more network neighbors a brand has, the more heterogeneous is the semantic context in which the brand name is used. This measure is higher when brand associations are more diverse and is consistent with previous research showing the positive effect of a higher number of associations on brand strength (Grohs, Raies, Koll, & Mühlbacher, 2016). The third component, connectivity, expresses how often a word (in our case a brand) serves as an indirect link between all the other pairs of words, while constructing a co-occurrence network (see Section 3). It reflects the embeddedness of a museum name in a discourse and can be considered as the expression of the connective power of a brand name, i.e. the ability to indirectly link different words and/or topics. This dimension is consistent with other studies. For example, Gloor et al. (2009) mapped semantic networks extracted from the web and found that the betweenness centrality of a brand could be used as a proxy for its popularity. Similarly, another study found that brand relevance in specific contexts can be measured via its betweenness centrality (Fronzetti Colladon & Scettri, 2019). While a brand name could be frequently mentioned (high prevalence) and might have heterogeneous associations to other brands or concepts (high diversity), the museum name could still be peripheral and not connected to the core of the conversations.

Overall, these considerations suggest that a museum, whose brand is frequently used in online forums (prevalence), that is embedded in a rich discourse (diverse), and is at the core of a discourse (connected), has a greater competitive advantage over other museums with a lower brand importance, in terms of the ability to attract customers to their site. When users perceive that a brand is supported and mentioned by other users, who offer detailed explanations of the brand value, or add a comparative analysis, they are more likely to be persuaded and they might follow a similar path. We therefore formulate the first hypothesis as follows:

H1. *There is a positive association between the importance of museum brands in online conversations, measured through the Semantic Brand Score, and the number of museum visitors.*

An important brand is at the core of a conversation, with the possibility of being associated to either negative or positive feelings. On the other side, a brand that is used marginally, or that is very peripheral in a set of documents, can be classified as unimportant. When positive words are used to talk about a brand online (i.e. positive sentiment), word of mouth will likely lead to an increase of visitors for the museum. Therefore, a positive sentiment of the words which co-occur with a museum name should reinforce and complement its brand importance.

As suggested by different authors (Dellarocas, Awad, & Zhang, 2004; Smith, Menon, & Sivakumar, 2005), online reviews can be perceived as more credible than traditional word-of-mouth as they are originated by users with similar attitudes and preferences. The more users talk about a brand in positive terms, the higher the chance for the online word of mouth to be translated in economic value for the brand/museum. In their process of searching for information and validation and select their final tourism destination, users might change their initial choice based on the positive words associated with a brand that they find on the forum. Secondly, when users leave a positive review or a positive rating for a product, which seems to happen more often than leaving a negative comment (Bilro, Loureiro, & Guerreiro, 2019), other users will be more likely to buy that product or like that brand (Hollebeek, Glynn, & Brodie, 2014).

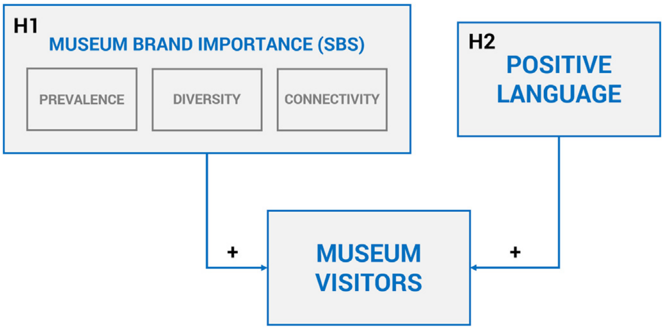


Fig. 1. Conceptual framework.

These considerations would lead us to assume that an increase in the positive sentiment associated with museum brands benefit the institutions by bringing in more visitors attracted by the peer-review validation. We therefore propose the following hypothesis:

H2. *The more positive the sentiment of online posts about a museum, the higher the number of visitors.*

Fig. 1 illustrates the conceptual framework and hypotheses driving this study:

3. Research design

3.1. Data collection

In this study, we considered the last 10 years of the online conversations happening on popular online forums, accessible also to non-registered users, where people exchange travel tips and opinions and share personal experiences on places and attractions. Travelers search online to increase the quality of future trips and to minimize potential risks associated with future travel (Jacobsen & Munar, 2012). A recent Nielsen report on social media (Shannon, Andrew, & Maeve, 2016) claims that travel websites are the second most-trusted source of brand information after recommendations from friends and family members.

For our analysis, we used an existing database comprising a large number of forum posts (Innarella, 2018). This database contains data on a selection of European museums: the Louvre and the Pompidou Centre in Paris, the Borghese Gallery and the Vittoriano in Rome, and the National Gallery in Prague. The data is made of more than 2,830,000 forum posts about tourist attractions in Paris, Rome and Prague, written by more than 113,700 users over ten years, from January 2007 to December 2016. Based on this existing database, we built a uniform naming conventions given the use of abbreviations for museums and attractions, and we corrected typing errors on some of the posts. The author's choice of selecting museums from the capital cities of France, Italy and Hungary was mostly experimental and explorative, with the aim of studying cities with a different attractive power, different attractions and characteristics. As reported by the United Nations World Tourism Organization, France and Italy were respectively first and fifth in terms of international arrivals, while Hungary increased its yearly international arrivals by 7% (UNWTO Tourism Highlights*: 2017 Edition*, 2017). A brief description of the museums included in the sample and the number of posts collected is presented in Table 1.

Information about museum visitors was collected considering the annual reports published on the website of each museum, as well as consulting the cultural aggregators Statistica Beniculturali and Egmus (http://www.statistica.beniculturali.it; http://www.egmus.eu). Since data was available on an annual basis, our dependent variable consists of 50 observations (5 museums × 10 years). To build data consistency, we calculated our predictors, sentiment and the Semantic Brand Score, with an annual frequency (e.g. the brand importance of Louvre in

Table 1
Museums included in the sample.

Museums	Description
Louvre, Paris	The world's largest art museum and a historic monument in Paris, France. In 2017, the Louvre was the world's most visited art museum, receiving 8.1 million visitors.
Borghese gallery, Rome	Art gallery in Rome, Italy, housed in the former *Villa Borghese Pinciana*. It houses a large part of the Borghese collection of paintings, sculpture and antiquities; it has twenty in rooms across two floors.
National gallery, Prague	The most important gallery in the Czech Republic with the largest collection of Czech and international art. The collections are presented in a number of historic structures within the city of Prague, as well as other places.
Vittoriano, Rome	The *Altare della Patria* (Altar of the Fatherland), also known as *Il Vittoriano*, is a white marble monument located in Rome, Italy, built in honor of Victor Emmanuel, the first king of a unified Italy.
Pompidou centre, Paris	A complex building designed by Renzo Piano and Richard Rogers, is home to the National Museum of Modern Art in Paris and is internationally renowned for its 20th and 21st century art collections.

2016). The database we accessed only included posts written in English. Limiting the analysis to one single language was important in our study, in order to be consistent in the measurement of semantic variables. Although the calculation of both sentiment and Semantic Brand Score can be adapted to multiple languages, it would be inappropriate to directly compare scores coming from posts written in different languages. In future replications of this study, we aim to see whether standardization would be sufficient to address this issue, or whether alternative approaches are feasible. Moreover, English was the most frequently used language by tourists of different nationalities (Innarella, 2018). In order to appropriately measure brand importance, we analyzed all posts about Paris, Rome and Prague, even those not including museum names. This is particularly relevant for measures such as connectivity, which require assessingthe position of a museum name in the co-occurrence network with respect to the general discourse.

3.2. Study variables

We pre-processed text data in order to remove stop-words (i.e. those words which usually provide little contribution to the meaning of a sentence, such as the word 'and'), punctuation and special characters. We changed every word to lowercase and extracted stems by removing word affixes (Jivani, 2011), by using the NLTK Snowball Stemmer algorithm (Perkins, 2014). To conduct these preliminary operations and to calculate the SBS indicator, we adopted the programming language Python. The most important libraries we used for network analysis task were NLTK (Perkins, 2014), for Natural Language Processing, and Graph-Tool (Peixoto, 2014).

The subsequent step was to transform text documents into a social network where nodes are words that appear in the text. An arc exists between a pair of nodes if their corresponding words co-occurred at least once; arc weights are determined by the frequency of co-occurrence. Following this procedure, we obtained 30 networks: 3 city forums (Paris, Rome and Prague), 10 years of conversations. In order to filter out negligible or less frequent co-occurrences, we retained only the arcs which had a minimum weight of 5. Based on methods used by previous studies (Fronzetti Colladon, 2018), we adopted a five-word window for the determination of co-occurrences maximum range.

Prevalence was calculated as the frequency by which a museum name was mentioned in the forum posts. Diversity is a measure of the heterogeneity of textual brand associations and is higher when brand associations are more diverse. Diversity has been operationalized through the degree centrality measure (Wasserman & Faust, 1994):

$$Diversity\,(museum_i) = d\,(g_i)$$

It corresponds to the degree of the node g_i which represents the museum brand: $d(g_i)$.

Connectivity reflects the 'brokerage power' of a museum name in the discourse about city attractions (Fronzetti Colladon, 2018). While a brand name could be frequently mentioned (high prevalence) and might have heterogeneous associations to other words (high diversity), the museum name could still be peripheral in the conversations. Connectivity, calculated as the betweenness centrality of the brand term (Fronzetti Colladon, 2018; Wasserman & Faust, 1994), can be considered as the expression of its connective power, i.e. the ability to indirectly link different words or groups of words (sometimes seen as discourse topics):

$$Connectivity\,(museum_i) = \sum_{j<k} \frac{d_{jk}\,(g_i)}{d_{jk}}$$

with d_{jk} equals to the number of the shortest paths linking the generic pair of nodes g_j and g_k, and $d_{jk}(g_i)$ equal to the number of those paths which contain the museum brand node g_i. As suggested in a more recent work of Fronzetti Colladon (2019), in the computation of connectivity we considered the inverse of arc weights in the determination of shortest network paths, and therefore calculated weighted betweenness centrality using the algorithm proposed by Brandes (2001).

To compare measures derived from different networks (i.e. one for each year and for each museum), we standardized the values of prevalence, diversity and connectivity. Standardization was carried out subtracting the median to each individual score (of words in each network) and dividing it by the interquartile range. The Semantic Brand Score was subsequently calculated as the sum of the standardized values of its components (Fronzetti Colladon, 2018). According to this standardization procedure, SBS scores can either be positive or negative – based on the importance of a certain term, i.e. a museum name. If a term had a negative score, it means that its unstandardized value is below the median of the scores obtained by the other significant words in the discourse.

Each of the above mentioned measures was calculated as the variation with respect to the previous year. A first differencing of the variables permitted the elimination of time trends and produced stationary data. A first differencing was also applied to the dependent variable of our study, i.e. the yearly number of museum visitors.

Lastly, we measured the sentiment of museum brands, considering the valence of their textual associations, obtained from the polarity scores included in the SenticNet 4 dictionary (Cambria, Poria, Bajpai, & Schuller, 2016). As each association had its own strength – represented by the co-occurrence frequency – overall brand sentiment was calculated as the weighted average of association polarity. This measure has been subsequently rescaled in the range [0,1] with values below 0.5 representing a negative sentiment on average, and values above this threshold indicating a prevalence of positive associations. Other approaches for the calculation of sentiment are also possible, and could be tested in future research. For example, one could train an ad-hoc classifier, using supervised machine learning algorithms. Another alternative is to use the VADER lexicon included in the NLTK package, which seems to work well for texts extracted from social media (Hutto & Gilbert, 2014). We additionally tested this approach, without drawing significantly different conclusions from our models.

Table 2 presents the descriptive statistics for our variables. A first interesting result is that the average sentiment of the textual brand

Table 2
Descriptive statistics.

Variable	M	SD	Min	Max
Visitors	2,568,946	3,249,613	18,215	9,437,744
Prevalence	184.156	274.296	0	830
Diversity	7.276	9.083	−0.163	27.109
Connectivity	122.166	216.580	−0.002	757.846
SBS	313.598	494.676	0	1594.787
Sentiment	0.579	0.083	0.411	0.917

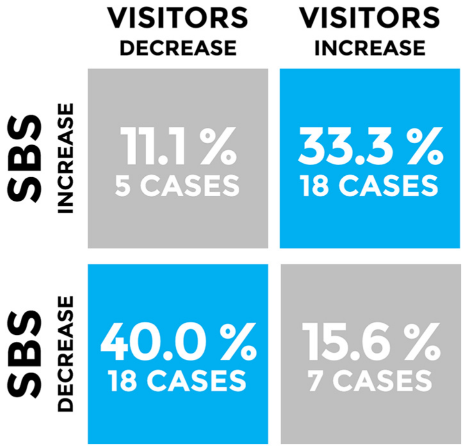
associations was positive, indicating that museums were mentioned on the forums with a usually positive valence of the words used. This might indicate that users have on average a positive attitude when they ask for or provide general information about museums (e.g. ticket price, location, opening hours, events and temporary exhibitions), with few posts expressing criticism. Standardized values of the SBS and its components are positive on average, suggesting that museum names got significant attention within the discourse about the three capital cities we analyzed.

4. Results

Our findings indicate that, as the SBS grows, so does the number of museum visitors, whereas a decrease in SBS is associated with a decrease of visitors. Fig. 2 represents a contingency table, which shows the concordance of yearly change of SBS with change in the number of museum visitors. Blue squares indicate concordant cases, whereas grey square discordant ones.

In about 73% percent of cases the signs of these variations are concordant, supporting the idea that an increase in the SBS can be indicative of a higher number of museum visitors, whereas a decrease in SBS often associates with a decrease of visitors. These results are also confirmed by the significant Pearson Chi-square test ($\chi^2 = 9.82$, $p = .002$) and by the Fisher's exact test ($p = .003$), both carried out on the contingency table of Fig. 2.

We subsequently built the multiple regression models presented in Table 3 to understand which variables could better explain change in museum visitors. In all the models, we controlled for the possible effect of time and included several dummy variables representing each museum considered in our study. This method was appropriate since the selected museums had a different average number of visitors per year and we had repeated measures over time. Another possibility to deal with repeated measures over time would be to use multilevel regression models (Hoffman & Rovine, 2007; Singer & Willett, 2003), as these models allow for an analysis of variance on multiple levels, to see which part is accountable for the differences in museums and which is residual. However, we built such models with respect to our dependent variable – nesting repeated measures (level 1) in museums (level 2) – and found that the intra-class correlation coefficient was close to zero, indicating a high dominance of the residual variance. In Model 1, we only included the control variables which we included in all models. In Models 2–5, we tested separately the three components of the SBS as well as the sentiment variable. Model 6 is the full model. In Models 7, we used the SBS instead of its separate components, and then combined it with sentiment in Model 8. Model 1 shows that control variables alone can explain about 17% of variance, however with a very low adjusted R-squared (0.05) and only Time is significant. In the subsequent models, we tested separately the contribution of each dimension of the SBS, finding that prevalence and connectivity are the most significant, together with the SBS aggregate measure, all presenting a positive coefficient. Prevalence seems to be the most important predictor, being able to improve the variance explained by Model 1 of about 26%, and its adjusted R-squared of about 33%. Combining the three dimensions in Model 6 led to the best results, with an R-squared of 57.55% and an adjusted R-squared of 45.23%. In this model, diversity becomes significant. We checked for multicollinearity problems and found no evidence to support them (maximum Variance Inflation Factor was 2.33 for Model 6). On the other hand, as the SBS is the sum of standardized prevalence, diversity and connectivity, collinearity problems would arise if putting in the same model the three dimensions together with the final indicator. Following the suggestion of Fronzetti Colladon (2018), we additionally explored the impact of the sentiment variable, which resulted always non-significant.



	VISITORS DECREASE	VISITORS INCREASE
SBS INCREASE	11.1 % 5 CASES	33.3 % 18 CASES
SBS DECREASE	40.0 % 18 CASES	15.6 % 7 CASES

Fig. 2. Association of visitors and SBS variations.

We additionally checked for the effects produced by time lags of our variables, always considering their first differencing. However, these did not lead to better models, also considering the limited number of observations in our sample (50 total, 10 per each of the 5 museums). This finding is consistent with previous work showing that future visitors usually consult the latest forum posts about the topic of their interest while older reviews are perceived as less informative (Wu, Che, Chan, & Lu, 2015); therefore, in most cases, posts written in the same year of their visit, not before. Older posts also rarely appear on the first page of a forum search query.

Lastly, since the National Gallery of Prague is probably less known than the other museums, we tried to remove it from the analysis, to check the robustness of our models. The new results were fully consistent with those of Table 3, with a slight improvement of the Adjusted R^2 (from 0.4523 to 0.4659, for the best model, Model 6).

Multiple regression was a first attempt to prove the significance and directionality of the association of brand importance with museum visitors. Using a Naïve Bayes algorithm (John & Langley, 1995), we extended the analysis, obtaining predictions of change in museum visitors which have a reasonable accuracy. We used the machine learning software Weka (Holmes, Donkin, & Witten, 1994) and – after testing many combinations of algorithms, including Random Forests and Support Vector Machine (Breiman, 2001; Suthaharan, 2016) – we found that the best results were those of Naive Bayes (John & Langley, 1995). This produced forecasts of positive/negative change in visitors which were 75.56% accurate and had the positive or negative variations of the SBS and its components as predictors. Our choice was also supported by the results obtained using the Auto-Weka package (Kotthoff, Thornton, Hoos, Hutter, & Leyton-Brown, 2017). A reasonable fit of the Naive Bayes algorithm was confirmed by the average values of the Cohen's Kappa and of the area under the ROC curve – respectively 0.51 and 0.75. The algorithm was trained on a random 70% of the sample, and the remaining 30% was left out for evaluation. The process was repeated 500 times and the accuracy values we reported are on average.

5. Discussion and conclusions

The recent availability of real-time, high-volume data, and the easy

Table 3
Regression models.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Time	−41,679.2*	−12,994.2	−42,565.3*	−32,454.58	−41,595.91*	−15,424.16	−25,979.1	−25,890.43
Louvre	−135,870	−36,534.6	−139,122.2	−143,816.2	−135,415	−65,345.1	−120,660.6	−120,183.7
National Gallery Prague	113,552.8	56,834.2	114,846	101,698.2	106,752.9	55,126.49	88,847.29	81,754.48
Complesso del Vittoriano Rome	−3297.6	−17,052.95	−2632.497	−2244.497	−89.539	−8716.255	−5480.918	−2136.305
Centre Pompidou	78,952.3	88,850.9	77,718.75	76,387.36	77,513.08	68,995.43	78,856.59	77,355.93
Prevalence		5129.8***				6758.791***		
Diversity			8115.49			142,279.4**		
Connectivity				1185.371*		511.5725		
SBS							1164.427**	1164.563**
Sentiment					87,153.26	16,047.88		90,871.11
Constant	254,434.8	96,251.36	259,252.1	198,028.2	203,777	96,802.04	162,426.9	109,597.4
R-Squared	0.1714	0.4317	0.1721	0.2853	0.1719	0.5755	0.3423	0.3428
Adj R-Squared	0.053	0.3314	0.026	0.1592	0.0257	0.4523	0.2262	0.2034

* $p < .05$.
** $p < .01$.
*** $p < .001$.

access to users' digital traces, offer a new opportunity to obtain improved understanding of tourist behaviors. This paper has described the application of a methodology based on the integration of social network analysis and text mining to measure brand importance and study its association with museum visitors: to this purpose we used the Semantic Brand Score (Fronzetti Colladon, 2018).

Our models support the first hypothesis and offer insights on the association between brand importance and change in museum visitors. The strongest proportion of variance explained in the regression models (57.6%) was obtained by combining the three dimensions of the SBS. Connectivity partially contributed to the increase of variance explained, but was not significant in the final regression model. However, connectivity, together with the other SBS dimensions, was important to improve the accuracy of the Naïve Bayes algorithm. The sentiment indicator, on the other hand, was never significant. It seems that when users provide articulated reviews on a museum – frequently mentioning its name, likely adding detailed explanations of pros and cons, and using heterogeneous words – the likelihood to convince others to finalize a purchase or reserve a museum pass is higher. This is aligned with a recent study on online consumer perception, review factuality and source credibility (Filieri, Hofacker, & Alguezaui, 2018), which suggests that consumers tend to look for reviews that report accounts of facts and events related to their experience. A study focused on Yelp.com comments provided similar evidences, showing that cognitive processing is more relevant than other components of brand affection and activation/energy (Bilro et al., 2019). Future tourists might be persuaded to follow other users' leads on a brand when information provided is more articulated, diverse and frequently discussed in several posts.

Our models suggest that what really matters in terms of building a strong and attractive brand is that users talk about the brand on social media, even if they provide comments that are not necessarily the most positive. Therefore, our second hypothesis was not supported by the models, since the sentiment of the words associated with a museum brand is not necessarily associated with an increase in its visitors over time. This contrasts with other studies indicating that purchasing behavior could be influenced by positive comments left by others (Kim & Ko, 2012). It seems that consumers tend to prefer posts that display rich information, rather than overly positive reviews, which is a result confirmed by other studies (Filieri et al., 2018; Filieri & McLeay, 2014).

The evidences collected in this study suggest that in order to increase museum visitors over time it is important to increase the volume of online posting and the richness of information generated by users around the brand. This seems to suggest that tourists might be influenced by the *awareness effect* generated by online word-of-mouth, that is the presence of brand names. In order to increase museum visitors over time, it is important to increase the volume of online posting, rather than controlling for the positivity or negativity of the posts. This is aligned with findings by Duan, Gu, and Whinston (2008) who found a positive association between online word-of-mouth and movie sales: whereas box office sales were significantly influenced by the volume of online posting, higher ratings did not lead to higher sales. The simple presence of brand reviews conveys the existence of the product which makes it more desirable by consumers (Godes & Mayzlin, 2004; Viglia, Minazzi, & Buhalis, 2016). Since the sentiment of the words associated to the brand was on average positive, we can imply that only a few users had expressed negative comments on the museums. This could mean that a positive sentiment on average is sufficient to attract new visitors and that being even more positive is not necessary to promote a positive brand image. It would be interesting to replicate this study in scenarios where sentiment variations are more pronounced.

Museums today perform their functions in an extremely competitive market environment, with some of them struggling to survive due to decreasing visitor numbers and financial bottlenecks in the public sector (Gretzel, Werthner, Koo, & Lamsfus, 2015; Kovaleva, Epstein, & Parik, 2018). Implementing management techniques typically adopted in the for profit sector and designing new brand management strategies has the potential to increase the likelihood of repeat visits and recommendations to visit through word of mouth. For the past two decades there has been an increasing interest in implementing marketing techniques in the museum context, which translated into a need to have museums become more marketing oriented (Rentschler, 2002; Viglia et al., 2016; Xiang et al., 2015). Tourism practitioners and arts administrators need accurate forecasts of tourist volume, in order to effectively allocate resources and formulate pricing strategies. Our paper provides empirical evidence that a methodology based on big data has the potential to help design and implement a branding policy to address negative trends among museums across the world (Ober-Heilig, Bekmeier-Feuerhahn, & Sikkenga, 2012). Traditional research used surveys and expensive observational studies to provide data to evaluate museum visitor behavior, with limits of scale and related bias. The main contribution of our research is to present a new big data approach to assess museum brand importance from the analysis of online forums, which can be correlated with data already available regarding museum visitors and their purchasing behaviors. This approach is based on the analysis of the discourse of a broad public and is less expensive than surveys. Our research offers additional evidences that can inspire researchers and practitioners in the tourism field to adopt big data methods for their decision-making processes. Museum brand managers could use metrics such as the ones included in the Semantic Brand Score to compare their brand's importance with competitor brands. They could analyze multiple sources of text data, such as social media or newspaper articles, and measure prevalence, connectivity and diversity. This study suggests to invest in marketing activities and resources that could improve online word of mouth, and increase the SBS of a brand. This means investing in a content marketing strategy that has the potential to increase the frequency with which the brand name appears within online documents (prevalence component of the SBS). Marketing managers should also carefully prepare detailed and rich content to increase the variety of information available to online users (diversity component of the SBS), as this appears to be associated with more visitors being attracted to the museum. In order to increase the connectivity of a brand within the overall conversation, managers could carefully design co-marketing strategies by partnering with institutions in the same geographic area (e.g. other museums, public sites, restaurants). The design of marketing synergies among institutions has the potential to increase the quality and depth of content offered to potential visitors, which our study shows to be associated to an increased chance of museum visits. Lastly, while it is important to monitor the sentiment of online users towards a brand, our results suggest that museum marketing managers should be less concerned about the positive or negative language used by tourists, and be more interested in improving the quality of the content provided. While most of the empirical studies thus far have been using social media and online forums to predict consumers' behaviors, the triangulation of user-generated data from various platforms represents an untapped potential. Besides looking at the total number of online comments their museum is receiving, administrators should closely review the diversity and connectivity of their brand, which our results suggest to be more impactful than sentiment.

This study extends the research on brand importance and the applications of the Semantic Brand Score that, to the extent of our knowledge, has never been used to evaluate museum brands or anticipate trends in museum visitors. Previous studies have assessed brand equity and brand importance via expensive and time consuming market surveys administered to consumers and other stakeholders, or via financial methods (Gretzel et al., 2015; Kovaleva et al., 2018). The approach we use, on the other hand, allows repeatable measurements, for a constant monitoring of brand importance with almost no additional cost. It is based on the analysis of big textual data, which come from the spontaneous conversations of tourists on online forums, without some of the biases induced by interviews (Pentland, 2010). Lastly, our findings partially contrast with studies attributing high importance to the positivity of messages for the prediction of purchasing behaviors (Kim & Ko, 2012), as sentiment in our setting was mostly uninfluential.

This study was based on a limited number of selected European museums. The sample size represents a limitation to the generalizability of the results; therefore, we recommend extending the application of this methodology to other European and non-European museums. A larger sample, where researchers could access more granular data – for example with a monthly frequency – could support the implementation of forecasting models or an in-depth time series analysis. For example, a larger dataset would allow a rolling window out-of-sample forecasting approach. Another limitation is intrinsic to all the methods based on quantitative textual analysis, since they cannot fully account for important factors impacting consumers' decisions, such as the perceived credibility of a source/reviewer or the currency of the review (i.e. how up to date is the information a reviewer is sharing). Future studies should explore the effects of other online user experience factors that might affect brand perception and economic outcome, including the official star-rating system, which is a key variable through which tourism destinations – such as museums, hotels, and restaurants – can differentiate their offering (Silva, 2015).

References

Aaker, D. A. (1996). Measuring brand equity across products and markets. *California Management Review, 38*(3), 102–120.

Aggarwal, P., Vaidyanathan, R., & Venkatesh, A. (2009). Using lexical semantic analysis to derive online brand positions: An application to retail marketing research. *Journal of Retailing, 85*(2), 145–158. https://doi.org/10.1016/j.jretai.2009.03.001.

Banerjee, S., & Chua, A. Y. K. (2016). In search of patterns among travellers' hotel ratings in TripAdvisor. *Tourism Management, 53*, 125–131. https://doi.org/10.1016/j.tourman.2015.09.020.

Barnes, S. J., Mattsson, J., & Sørensen, F. (2014). Destination brand experience and visitor behavior: Testing a scale in the tourism context. *Annals of Tourism Research, 48*, 121–139. https://doi.org/10.1016/j.annals.2014.06.002.

Belén del Río, A., Vázquez, R., & Iglesias, V. (2001). The effects of brand associations on consumer response. *Journal of Consumer Marketing, 18*(5), 410–425. https://doi.org/10.1108/07363760110398808.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. *Journal of Hospitality Marketing and Management*. https://doi.org/10.1080/19368623.2018.1506375.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bucur, C. (2015). Using opinion mining techniques in tourism. *Procedia Economics and Finance.*. https://doi.org/10.1016/s2212-5671(15)00471-2.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. *26th International Conference on Computational Linguistics (COLING)* (pp. 2666–2677). Osaka.

Capodieci, A., Elia, G., Grippa, F., & Mainetti, L. (2019). A network-based dashboard for cultural heritage promotion in digital environments. *International Journal of Entrepreneurship and Small Business*. https://doi.org/10.1504/IJESB.2019.098986.

De Mauro, A., Greco, M., Grimaldi, M., De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *Proceedings of the 4th International Conference on Integrated Information, 1644*(May), 97–104. https://doi.org/10.1063/1.4907823.

De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing, 26*(2), 83–91. https://doi.org/10.1016/j.intmar.2012.01.003.

Dellarocas, C., Awad, N. F., & Zhang, X. (2004). Exploring the value of online reviews to organizations : Implications for revenue forecasting and planning. *ICIS 2004* (pp. 379–386). Washington, DC: Association for Information Systems.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? - an empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007–1016. https://doi.org/10.1016/j.dss.2008.04.001.

Evert, S. (2005). *The statistics of word Cooccurrences word pairs and collocations*.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management, 52*, 498–506. https://doi.org/10.1016/j.tourman.2015.07.018.

Filieri, R., Hofacker, C. F., & Alguezaui, S. (2018). What makes information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2017.10.039.

Filieri, R., & McLeay, F. (2014). E-WOM and accommodation. *Journal of Travel Research*. https://doi.org/10.1177/0047287513481274.

Fronzetti Colladon, A. (2018). The semantic brand score. *Journal of Business Research, 88*, 150–160. https://doi.org/10.1016/j.jbusres.2018.03.026.

Fronzetti Colladon, A. (2019). Forecasting election results by studying brand importance in online news. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2019.05.013 In press.

Fronzetti Colladon, A., & Scettri, G. (2019). Look inside. Predicting stock prices by analysing an Enterprise intranet social network and using word co-occurrence networks. *International Journal of Entrepreneurship and Small Business, 36*(4), 378–391. https://doi.org/10.1504/IJESB.2019.10007839.

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management, 66*, 53–61. https://doi.org/10.1016/j.tourman.2017.10.018.

Gloor, P. A. (2017). *Swarm leadership and the collective mind: Using collaborative innovation networks to build a better business.* Bingley, UK: Emerald Publishing Limited.

Gloor, P. A., Krauss, J., Nann, S., Fischbach, K., & Schoder, D. (2009). Web science 2.0: Identifying trends through semantic social network analysis. *2009 international conference on computational science and engineering* (pp. 215–222). Vancouver, Canada: IEEE. https://doi.org/10.1109/CSE.2009.186.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science, 23*(4), 545–560. https://doi.org/10.1287/mksc.1040.0071.

Gretzel, U., Werthner, H., Koo, C., & Lamsfus, C. (2015). Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2015.03.043.

Grohs, R., Raies, K., Koll, O., & Mühlbacher, H. (2016). One pie, many recipes: Alternative paths to high brand strength. *Journal of Business Research, 69*(6), 2244–2251. https://doi.org/10.1016/j.jbusres.2015.12.037.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101–117.

Hollebeek, L. D., Glynn, M. S., & Brodie, R. J. (2014). Consumer brand engagement in social media: Conceptualization, scale development and validation. *Journal of Interactive Marketing, 28*(2), 149–165. https://doi.org/10.1016/j.intmar.2013.12.002.

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: a machine learning workbench. *Proceedings of ANZIIS '94 - Australian New Zealnd intelligent information systems conference* (pp. 357–361). . https://doi.org/10.1109/ANZIIS.1994.396988.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *KDD-2004 - proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*.

Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). The impact of user interactions in social media on brand awareness and purchase intention: The case of MINI on Facebook. *The Journal of Product and Brand Management, 22*(5/6), 342–351. https://doi.org/10.1108/JPBM-05-2013-0299.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the eighth international AAAI conference on weblogs and social media* (pp. 216–225). Ann Arbor, Michigan, USA: AAAI Press.

Hwang, Y. H., Jani, D., & Jeong, H. K. (2013). Analyzing international tourists' functional information needs: A comparative analysis of inquiries in an on-line travel forum. *Journal of Business Research, 66*(6), 700–705. https://doi.org/10.1016/j.jbusres.2011.09.006.

Innarella, R. (2018). *Tecniche di estrazione ed analisi di big data in ambito turistico,* 2.0.

Jacobsen, J. K. S., & Munar, A. M. (2012). Tourist information search and destination choice in a digital age. *Tourism Management Perspectives, 1*(1), 39–47. https://doi.org/10.1016/j.tmp.2011.12.005.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *International Journal of Computer Technology and Applications, 2*(6), 1930–1938 doi:10.1.1.642.7100.

John, G. H., & Langley, P. (1995). Estimating continuous distribution in Bayesian classifiers. *Eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). San Mateo, CA: Morgan Kaufmann Publishers.

Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing, 57*(1), 1–22.

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research, 65*(10), 1480–1486. https://doi.org/10.1016/j.jbusres.2011.10.014.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research, 18*(25), 1–5.

Kovaleva, A., Epstein, M., & Parik, I. (2018). National heritage branding: A case study of the Russian museum of ethnography. *Journal of Heritage Tourism, 13*(2), 128–142. https://doi.org/10.1080/1743873X.2017.1343337.

Lassar, W., Mittal, B., & Sharma, A. (1995). Measuring customer-based brand equity. *Journal of Consumer Marketing, 12*(4), 11–19. https://doi.org/10.1108/07363769510095270.

Lee, H. A., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing, 28*(7), 675–688. https://doi.org/10.1080/10548408.2011.611739.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22. https://doi.org/10.1080/10548408.2013.750919.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66. https://doi.org/10.1016/j.tourman.2016.07.005.

Mandal, S. (2018). Exploring the influence of big data analytics management capabilities on sustainable tourism supply chain performance: The moderating role of technology orientation. *Journal of Travel & Tourism Marketing, 35*(8), 1104–1118. https://doi.org/10.1080/10548408.2018.1476302.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information and Management, 54*(6), 771–785. https://doi.org/10.1016/j.im.2016.11.011.

Miguéns, J., Baggio, R., & Costa, C. (2008). Social media and tourism destinations: TripAdvisor case study. *Advances in Tourism Research, 26*(28), 26–28. https://doi.org/10.1088/1751-8113/44/8/085201.

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas strip. *Tourism Management Perspectives, 23*, 41–52. https://doi.org/10.1016/j.tmp.2017.04.003.

Munar, A. M., & Ooi, C. S. (2012). The truth of the crowds: Social media and the heritage experience. *The cultural moment in tourism* (pp. 255–273). . https://doi.org/10.4324/9780203831755.

Neirotti, P., Raguseo, E., & Paolucci, E. (2016). Are customers' reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning. *International Journal of Information Management, 36*(6partA), 1133–1143. https://doi.org/10.1016/j.ijinfomgt.2016.02.010.

Ober-Heilig, N., Bekmeier-Feuerhahn, S., & Sikkenga, J. (2012). How to attract visitors with strategic, value-based experience design. *Marketing ZFP, 34*(4), 301–315. https://doi.org/10.15358/0344-1369-2012-4-301.

Padhi, S. S., & Pati, R. K. (2017). Quantifying potential tourist behavior in choice of destination using Google trends. *Tourism Management Perspectives, 24*, 34–47. https://doi.org/10.1016/j.tmp.2017.07.001.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing, 34*(3), 357–368. https://doi.org/10.1080/10548408.2016.1170651.

Peixoto, T. P. (2014). *The graph-tool python library.* https://doi.org/10.6084/m9.figshare.1164194.

Pentland, A. (2010). *Honest signals: How they shape our world.* Cambridge, MA: MIT Press.

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook. Python 3 text processing with NLTK 3 cookbook.* Birmingham, UK: Packt Publishing.

Rentschler, R. (2002). Museum and performing arts marketing: The age of discovery. *The Journal of Arts Management, Law, and Society, 32*(1), 7–14. https://doi.org/10.1080/10632920209597330.

Shannon, G., Andrew, P., & Maeve, D. (2016). *Demographics of social media users in 2016*.

Silva, R. (2015). Multimarket contact, differentiation, and prices of chain hotels. *Tourism Management, 48*, 305–315. https://doi.org/10.1016/j.tourman.2014.11.006.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Presshttps://doi.org/10.1093/acprof:oso/9780195152968.001.0001.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing, 19*(3), 15–37. https://doi.org/10.1002/dir.20041.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323. https://doi.org/10.1016/j.tourman.2010.12.011.

Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification* (pp. 207–235). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-7641-3_9.

Torres, E. N., Singh, D., & Robertson-Ring, A. (2015). Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry. *International Journal of Hospitality Management, 50*, 77–83. https://doi.org/10.1016/j.ijhm.2015.07.012.

UNWTO (2017). *Tourism highlights: 2017 Edition.* World Tourism Organization UNWTO.

Viglia, G., Minazzi, R., & Buhalis, D. (2016). The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-05-2015-0238.

Volcheck, E., Song, H., Law, R., & Buhalis, D. (2018). Forecasting London museum visitors using Google trends data. *e-Review of tourism research* (pp. 1–5). Sweden: Jönköping.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* New York, NY: Cambridge University Presshttps://doi.org/10.1525/ae.1997.24.1.219.

Wong, C. U. I., & Qi, S. (2017). Tracking the evolution of a destination's image by text-mining online reviews - the case of Macau. *Tourism Management Perspectives, 23*, 19–29. https://doi.org/10.1016/j.tmp.2017.03.009.

Wood, L. (2000). Brands and brand equity: Definition and management. *Management Decision, 38*(9), 662–669. https://doi.org/10.1108/00251740010379100.

Wu, C., Che, H., Chan, T. Y., & Lu, X. (2015). The economic value of online reviews. *Marketing Science, 34*(5), 739–754. https://doi.org/10.1287/mksc.2015.0926.

Xiang, Z. (2018). From digitization to the age of acceleration: On information technology and tourism. *Tourism Management Perspectives, 25*, 147–150. https://doi.org/10.1016/j.tmp.2017.11.023.

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 27*(2), 634–639. https://doi.org/10.1016/j.chb.2010.04.014.

Yun, Q., & Gloor, P. A. (2015). The web mirrors value in the real world: Comparing a firm's valuation with its web network position. *Computational & Mathematical Organization Theory, 21*(4), 356–379. https://doi.org/10.1007/s10588-015-9189-6.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear.". *Procedia - Social and Behavioral Sciences, 26*, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562.

Andrea Fronzetti Colladon, Ph.D., is Assistant Professor at the University of Perugia. He has been Research Fellow and Adjunct Professor of Engineering Economics at the University of Rome Tor Vergata, and Visiting Ph.D. Student at the MIT Center for Collective Intelligence – where he now collaborates on several research projects. Dr. Fronzetti Colladon is member of the ICKN core team and instructor of four courses on Social Network and Big Data Analysis. His research and scholarship interests include social network analysis, text mining, innovation management, organizational communication and brand management.

Francesca Grippa, Ph.D., is Full Teaching Professor and Faculty Director at Northeastern University. Dr Grippa's research interests include: collaborative innovation networks, entrepreneurship and change management. Dr Grippa is member of a research project at the MIT Center for Collective Intelligence that focuses on the application of dynamic network analysis to investigate the diffusion of innovation. She obtained a PhD in e-Business Management from University of Salento, Italy, and was a visiting scholar at the MIT Center for Digital Business.

Rosy Innarella received her Ph.D. in Enterprise Engineering from the University of Rome "Tor Vergata", Italy, in 2018. Her research interests include social network and semantic analysis, big data analytics, smart urban transport, smart tourism.