

## Research Article

# Business Analytics in Tourism: Uncovering Knowledge from Crowds

Carla Marcolin<sup>1</sup>  
João Luiz Becker<sup>2</sup>  
Fridolin Wild<sup>3</sup>  
Giovana Schiavi<sup>4</sup>  
Ariel Behr<sup>4</sup>

Universidade Federal de Uberlândia, Uberlândia, MG, Brazil<sup>1</sup>  
Fundação Getulio Vargas, São Paulo, SP, Brazil<sup>2</sup>  
Oxford Brookes University, Oxford, UK<sup>3</sup>  
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil<sup>4</sup>

Received 20 November 2018. This paper was with the authors for three revisions. Accepted 25 June 2019. First published online 19 July 2019.

Luiz Augusto Machado Mendes Filho was the associate editor for this article.

Editorial assistant: Luciane Kato Kiwara

Editor-in-chief: Carlo Gabriel Porto Bellini

## Abstract

Business Analytics leverages value from data, thus being an important tool for the decision-making process. However, the presence of data in different formats is a new challenge for analysis. Textual data has been drawing organizational attention as thousands of people express themselves daily in text, like the description of customer perceptions in the tourism and hospitality area. Despite the relevance of customer data in textual format to support decision making of hotel managers, its use is still modest, given the difficulty of analyzing and interpreting the large amounts of data. Our objective is to identify the main evaluation topics presented in online guest reviews and reveal changes throughout the years. We worked with 23,229 hotel reviews collected from TripAdvisor website through WebScrapping packages in R, and used a text mining approach (Latent Semantic Analysis) to analyze the data. This contributes with practical implications to hotel managers by demonstrating the applicability of text data and tools based on open-source solutions and by providing insights about the data and assisting in the decision-making process. This article also contributes in presenting a stepwise text analysis, including capturing, cleaning and formatting publicly available data for organizational specialists.

**Keywords:** text mining; business analytics; hotel reviews.

## Introduction

The decision-making process, in different managerial environments, faces many challenges as increasingly larger pools of data are produced every day. Companies are progressively pressed to access these data using analytics tools to support their decisions (Ransbotham & Kiron, 2017). Indeed, the concept and practice of Business Analytics had significant growth in the last decade, attracting the attention of researchers and managers from different areas (Mortenson, Doherty, & Robinson, 2015).

Business Analytics allows leveraging value from data, thus being an important tool for the decision-making process (Acito & Khatri, 2014). Business Analytics helps the analysis of large amounts of data and integrates different data sources, making it possible to improve a company's performance and identify business opportunities (Bayrak, 2015). However, in recent years, the presence of data in different formats poses extra challenges. Companies, in addition to dealing with large volumes of data, now need to handle data types such as voice, text, log files, images and videos (Davenport & Dyché, 2013).

In such a diverse context, textual data has been drawing organizational attention, as millions of people express themselves daily using text in many applications and tools available. If appropriately processed, textual data represent a perception sensor about customer experiences that is not only useful but vital for business analysis and decisions (Zhao, 2013). In this sense, the tourism and hospitality industry has been interested in customer perceptions for improving operations in the services industry (Han, Mankad, Gavirneni, & Verma, 2016). Nowadays, travel platforms like TripAdvisor collect volunteered information, openly distributing user reviews to firms and managers, challenging them with a great volume of data (Yoo, Sigala, & Gretzel, 2016) and building reputational economies for the tourism and hospitality industry (Langley & Leyshon, 2017).

Since platforms that facilitate experience sharing have become more and more popular, customers are willing to rely on electronic word-of-mouth (eWOM) as an important step before making a destination decision (Sparks & Browning, 2011). As most data are available in text format, finding effective ways to analyze and transform them into valuable information is one of the challenges that connects this industry to Business Analytics (Tang & Guo, 2015). As eWOM provides genuine information about customers, their opinions about tourism and hospitality services, expressed in natural language, form an important source of information for hotel managers (Carrasco & Villar, 2012).

However, despite the relevance of customer data in textual format to support the decision making of hotel managers, its use is not as frequent as it would be expected, due to the difficulty of analyzing and interpreting large amounts of data in that format, making it hard to acquire useful information for hotel strategy (He et al., 2017). In this way, capturing an accurate and complete picture of the customer experience is a most challenging task for hotel managers in recent years (Han et al., 2016).

In addition, there is evidence of concern about how to develop strategies to respond to such issues, given that, for most companies in the tourism and hospitality industry, there has been a change in managerial logic, pushed by the market and with intense use of eWOM platforms (Del Vecchio, Mele, Ndou, & Secundo, 2017). Given that online content is being produced faster than the capacity to analyze it (Ferreira, 2019) and developing strategies to adequately respond to customer's needs is an urgent need in this industry (Del Vecchio et al., 2017), the present article aims to answer the following research question: how can hotel managers analyze a large volume of data on guest reviews, so as to gain information and develop strategic actions in line with market trends? More specifically, our main objective is to identify the key evaluation topics presented in online guest reviews, revealing growing or falling trends through the years. This contributes to practices of hotel managers by developing and demonstrating the applicability of text mining tools, based on open-source solutions, and providing insights from the data and assisting in their strategic decision-making process. In addition, this article contributes with theory by demonstrating how to combine unsupervised learning and longitudinal analysis to make market trends evident, using publicly available customer textual data.

Although the evidence supporting the importance of reviews for hotel managers has already been explored, our approach is different from previous research in three aspects. First, rather than conducting experiments or focus group sessions (Horner & Swarbrooke, 2016; Sparks & Browning, 2011), this study analyzes real-world data extracted directly from websites that provide open access to information, like TripAdvisor. The data come from texts written directly by customers, representing the Voice of the Customer (VOC) itself and making it possible to understand what the customers are sharing about the organizations (Spangler & Kreulen, 2007). Second, the use of unsupervised learning tools such as Latent Semantic Analysis (LSA) allows for an objective analysis (Ashton, Evangelopoulos, & Prybutok, 2014), since the emergent categories are not provided by the analyst or taken from platforms (Xu, 2018), neither taken from any keywords framework or pre-existing ontologies (Thomaz, Bizb, Bettonic, Mendes-Filho, & Buhalise, 2017), rather emerging from the text, given the latent semantic relation between reviews and words. Finally, in order to provide strategically useful information, we analyzed topic trends through the years, allowing for the identification of growing or falling aspects of interest in the customer's view, instead of delivering a photograph of an instant in time, as usually done in other studies (Xu, 2018; Xu, Wang, Li, & Haghighi, 2017).

The article is organized as follows: in second section, we discuss some aspects related to the customer's review presented in the literature, as well as descriptions of the model applied; in third section, we present the methodological procedures adopted in the study; in fourth section, we discuss the results; and, in fifth section, we present final remarks and conclusions.

## Conceptual Background

As this work identifies the main topics of online guest reviews by revealing the evolution throughout the years, we explored textual data in a Business Analytics framework in the tourism and hospitality industry. Our conceptual background includes several papers that have explored guest reviews regarding the decision-making process of choosing hotels or destinations, thus

demonstrating the power of eWOM and review ratings in social media. Additionally, we analyze previous works on LSA, which was chosen to demonstrate the value of text analysis in decision support.

## Guest reviews

Accommodation and hotel services have high impact in tourism development (Vieira, Hoffmann, & Alberton, 2018). From the customer perspective, the importance of previous reviews for their decision-making process regarding hospitality has been extensively demonstrated in the literature (Sparks & Browning, 2011; Ye, Law, & Gu, 2009). Even without knowing the other users behind the screen, one important step in planning a travel, and thus deciding a place to stay, is to access a review from well-known websites and take that information in consideration. Social media and customer review websites, like TripAdvisor, have changed the tourism and hospitality industry and the practices of hotel managers (Molinillo, Sandoval, Morales, & Stefaniak, 2016). In the tourism literature, studies about eWOM have developed quickly over the last years with the increased popularity of customers' online booking and online review behavior (Xu, 2018).

Another important aspect is the strong predictive power of the so-called social media review rating and hotel performance metrics. Kim and Park (2017) compared traditional customer satisfaction of a hotel with the same data from four different websites. They discovered that not only social media ratings were better predictors for metrics like average daily rate and percentage of occupancy, but also that data from TripAdvisor had the closest correlation. Thus, social media rating is a significant predictor to explain hotel performance metrics like percentage of occupancy and room revenue (Kim & Park, 2017). Besides that, eWOM is associated with customer retention and loyalty, as online reputation comparison is facilitated through travel platforms (Cantalops & Salvi, 2014).

Previous experience from other customers has high importance before booking a hotel room online. Positive online reviews can significantly increase hotels booking rates. Besides, the polarity of reviews has a negative impact on reservations. Indeed, the tourism and hospitality industry should strongly consider online reviews, especially those posted in external portals apart from the organization's website (Ye et al., 2009). The review itself also tends to have more importance for customer perception, conveying more impact than ratings alone (Sparks & Browning, 2011).

Yen and Tang (2015) analyzed the motivations for posting hotel experiences with the online media chosen and identified whose eWOM motivations are affected by hotel attribute performance. The choice between TripAdvisor and Facebook, for example, is correlated with different motivations. TripAdvisor is associated with altruism and platform assistance, while Facebook is positively associated with extraversion, social benefits, and dissonance reduction. The findings suggest that motivations are not universally equal and eWOM behaviors are correlated with different motivations.

In this sense, recent advances in computer science, especially in Natural Language Processing (NLP), make it possible to work not only with ratings and other metrics, but also with text. Text

has a stronger power regarding customer decision (Lee, Jeong, & Lee, 2017), thus it should be included in the analysis agenda of hotel managers. Research by Perez-Aranda, Anaya-Sanchez, and Ruizalba (2017) explores this issue in a survey with 301 hotel managers. The main results show that managers are concerned with this type of platform, revealing the importance they place in analyzing customer opinions.

Furthermore, Han, Mankad, Gavirneni and Verma (2016) claim that hotel ratings do not tell the full story of how guests view a hotel. They found that negative comments have more weight in a guest's ratings of a hotel than positive comments. Such uneven weighting means that a simple average of positive and negative scores may not provide a clear view of guests' opinions. The study applied a regression analysis to the relationships of 18,106 distinct terms relating to five specific attributes: amenities, experience, location, transactions, and value. Each attribute was analyzed and provided important information to help hotel managers in their decision-making process.

Having a strong predictive power and being an important element for customer decision-making, those evidences reinforce the importance, for hospitality practitioners, to analyze objectively this type of data. A careful analysis can help managers to better understand what potential customers will face while searching for options. This article aims to help in this task.

## **Beyond text: LSA**

Textual data processing is not new, although the emergency of Business Analytics has placed it in highlights. Disciplines like information science have addressed issues of textual data indexing and organizing for quite a long time. More recently, advances in computer science tools have supported specific techniques and models in such tasks, as well as in information retrieval (IR) and document relevance (Manning, Rhagavan, & Schutze, 2009). At the same time, data mining gears have been increasingly modernized to meet the large volume of existing data, and the field has worked on how to solve manipulation and analysis issues in order to keep up with the increasing dynamics and speed of processes (Aggarwal & Zhai, 2012).

In this context, LSA is one of the models developed in response to the different needs of the IR area. More recently it has supported text mining activities (Visinescu & Evangelopoulos, 2014). When proposed by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), its main objective was to analyze synonymy and polysemy, working with texts in unstructured format. The authors looked for a tool that could recover more relevant documents by focusing on compatibility issues between the terms.

The purpose was to address the fact that it is unwise to use only term frequency or raw data for text indexing. Since textual data is available directly from people (thus, it is straightforward), it is important to consider that there are different ways to communicate a concept. Therefore, the terms used by a particular person may not match the terms used by another person, although they may express the same idea. On the other hand, two individuals may choose the same word to express different opinions. Such problems, namely synonymy and polysemy, are LSA's main concern (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

In addition, being an unsupervised learning technique, LSA conveys consistency and objectivity to the analysis, thus helping to avoid bias that could arise from other text analysis techniques. LSA model works with the concept of bag-of-words, that is, the sequence of terms and their context in the text is not considered. Each document in a collection of  $n$  documents (corpus) to be analyzed is formed by a set of words, called terms. By collecting all terms from all documents we form a set of  $m$  terms (bag-of-words). The presence or absence of a term in a document properly weighted forms an  $m \times n$  Term-Document Matrix.

The weighting scheme adopted in this paper is the index most commonly used in text mining, tf-idf. The first part is the term frequency (tf), that relates the frequency of a term  $t$  in a document  $d$  ( $f_{t,d}$ ) with the higher-frequency term in the same document ( $\max_{t' \in d} f_{t',d}$ ), that is,

$$\text{tf}_{t,d} = \frac{f_{t,d}}{\max_{t' \in d} f_{t',d}}.$$

The second part is the inverse of document frequency (idf), that is, the logarithm of the ratio between the number of documents in the collection being analyzed ( $n$ ) and the number of documents in the collection that contain a given term  $t$  ( $df_t$ ). More specifically,

$$\text{idf}_t = \ln \frac{n}{df_t}.$$

Both weights are then combined to produce a composite weight (tf-idf), that is,

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t,$$

which are the elements of the Term-Document Matrix. These entries value rare terms, which have the power to distinguish documents, and that, when present, have a significant frequency comparing to other terms in the same document. In addition, this index devalues terms that occur in many documents, penalizing them even with the value 0 when they are present in all documents (Crain, Zhou, Yang, & Zha, 2012).

This weighting scheme emerged from understanding the limitation of the Boolean model (0 and 1) that tends to oversimplify information from a set of documents by only considering the presence or absence of a term in a document. In the end, documents are represented as vectors of term weights in a vector space, allowing the application of concepts such as measures, distances and similarities between documents (Baeza-Yates & Ribeiro-Neto, 2011).

In order to discover topics concealed in the documents, LSA uses singular-value decomposition (SVD) of the Term-Document Matrix. This mathematical operation allows for the discovery of a latent semantic structure hidden among the terms from a set of documents. SVD factors any real matrix  $X$  of order  $m \times n$  as a product of three other matrices,

$$X = USV'.$$

The first matrix,  $U$ , is formed by the left singular vectors of  $X$  (normalized eigenvectors of  $XX'$  corresponding to its non-zero eigenvalues); the second one,  $S$ , is a diagonal matrix formed by the non-zero singular values of  $X$  (positive square roots of the non-zero eigenvalues of  $XX'$ ); and the third one,  $V'$ , is the transpose of the matrix formed by the right singular vectors of  $X$  (normalized eigenvectors of  $X'X$  corresponding to its non-zero eigenvalues). The order of  $S$  corresponds to the number of non-zero singular values of  $X$ , say,  $r$ .  $U$  is, therefore, of order  $m \times r$ , and  $V$  is of order  $n \times r$ . In addition, the columns of  $U$ , being eigenvectors of  $XX'$ , are orthogonal to each other. As they are normalized, they are of length 1. Similarly, the columns of  $V$  are also unitary vectors orthogonal to each other. We have, then,

$$U'U = V'V = I_r.$$

The columns of  $U$  form a linearly independent set, and therefore serve as a basis for the vector space generated by the columns of  $X$  (corresponding to documents in the corpus), being able to form any other vector in this same space from a linear combination of its elements. Similarly, the set of columns of  $V$  is linearly independent, forming a basis for the vector space generated by the rows of  $X$  (corresponding to terms in the bag-of-words), and therefore any vector in this space can be represented by a linear combination of its elements. Such orthonormal transformations preserve some properties of the original matrix  $X$ , like the length and distance of its row and column vectors (Martin & Berry, 2011).

When working with textual data, it is common to have a sparse Term Document Matrix, since, usually, different words will be presented in different documents with low frequency. Yet, there are usually many redundancies in the matrix, thus retaining only a small number of non-zero singular values loses little information. For that reason, we work with the first larger  $k$  singular values, reducing the dimensionality of the representation, leading to the main LSA equation, based on SVD:

$$X \cong U_k S_k V_k^t,$$

in which  $U_k$  and  $V_k$  are  $m \times k$  and  $n \times k$  matrices formed respectively by the normalized left and right eigenvectors of  $X$  corresponding to the  $k$  largest eigenvalues of  $XX'$  (or of  $X'X$ ) and  $S_k$  is a diagonal matrix formed by the  $k$  largest singular values of  $X$ .

Previous works with LSA include studies not only in computer science and statistics, but also in other applied areas, like education and marketing. In education, researchers are looking for task automatization, such as automatic correction (known as essay grading) or automatic feedback for students (Olmos, Jorge-Botana, Luzón, Martín-Cordero, & León, 2016). There are also studies investigating main concepts in textbooks and handouts in specific domains and their relationship with student development (Tinkler & Woods, 2013). Other studies show how the discovered topics can help tutors of large classes in distance learning models to better understand the students and help them more accurately (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999).



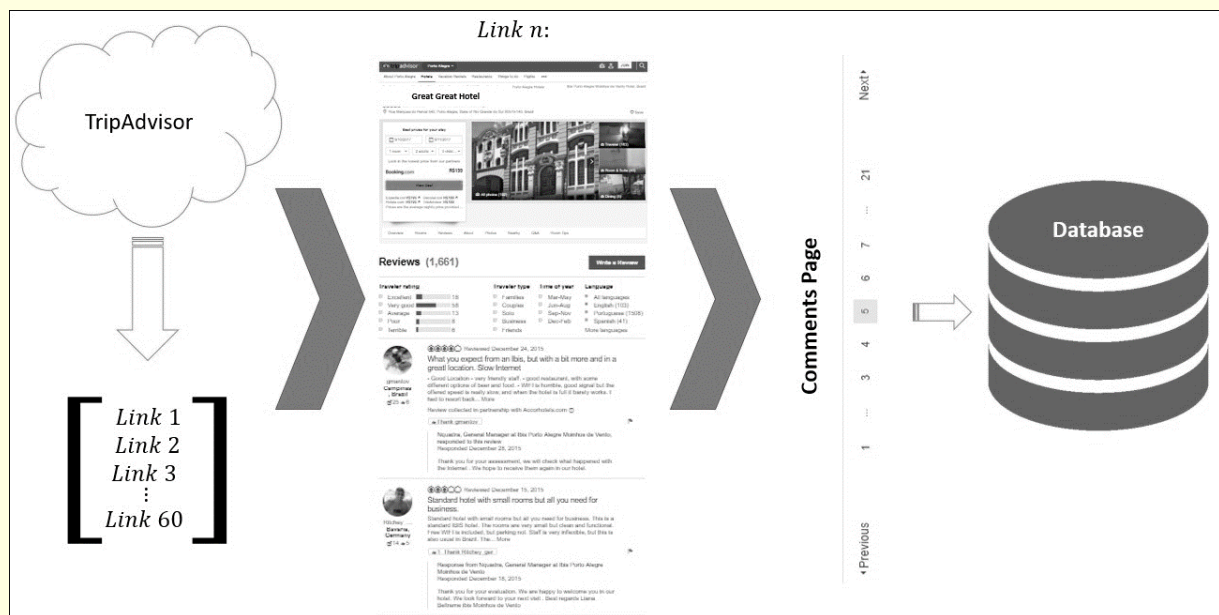
In marketing, exploring material available on the Web either by companies or customers, is the main objective. Thorleuchter and Van Den Poel (2012) showed that textual information from the website of e-commerce companies might be related to their success. We also found a study in the area of human resources that investigates the main skills demanded in job offers (O’Leary, Lindholm, Whitford, & Freeman, 2002). In a recent study, Turrell, Speigner, Djumalieva, Copple, and Thurgood (2019) use a dataset of 15 million UK job advertisements from a recruitment website to develop new economic statistics measuring labor market demand.

## Database and Algorithms

Consumer-to-consumer e-recommendation is growing. Also known as eWOM, tools to work with it are expanding, with more and more customers not only reading but also sharing their previous experiences, which are used as input before choosing a process, products or services (Tang & Guo, 2015). A well-known eWOM website in the tourism and hospitality industry is TripAdvisor, which we use as the data source for the study. TripAdvisor is a trend shaper regarding customer behavior in tourism and considered very reliable (Horner & Swarbrooke, 2016).

Such website can provide useful information for both companies and customers. Sometimes, a customer cannot read multiple comments, and a single bad or good experience reported may not be enough to take a properly informed decision, which would require reading more comments in an exhausting process. Likewise, any managerial conclusion about a service or a product should not be based on a small number of comments. Thus, for a robust analysis that can support decision making, the same process of reading and analyzing everything that has already been written is required, and that would be costly if done manually. We thus propose a novel method based on open-source tools to identify the main evaluation topics presented in online guest reviews, by extracting, treating and analyzing the main topics from a large set of documents.

Hotel reviews from TripAdvisor’s website were collected through WebScraping packages in R. With this tool, and based on Wickham (2015) previous development, it was possible to build an automated routine that collected complete comments from a specific city, using as input the complete link from each of the hotels listed in TripAdvisor (Figure 1). Being a function based on TripAdvisor structure, it can be used for any other listed city.



**Figure 1.** WebScraping process

To increase the representativeness of the sample, we considered all hotels located in the city of Porto Alegre with at least 100 comments registered in TripAdvisor. Porto Alegre is the southernmost Brazilian state capital and one of the 65 key tourism destinations selected by the Brazilian Tourism Ministry (Ministério do Turismo [MTur]) and integrating the Brazilian Competitiveness Model (Modelo de Competitividade Brasileiro [BCM]) (Instituto Brasileiro de Geografia e Estatística [IBGE], 2008; Vieira et al., 2018). In addition, Porto Alegre was considered an “efficient destination”, receiving large investments in marketing and promotion from BCM (Vieira et al., 2018, p. 908). The 100-comment limit was set after considering numerical relevance when building the mathematical structures after empirical tests.

Although destination management is an important factor for developed and developing countries, a common limitation in destination competitiveness models is the lack of indicators with empirical application to analyze and compare single destinations (Vieira et al., 2018) that may vary considerably within a country, especially in continental nations like Brazil. By considering a significant amount of reviews from a single destination, it is possible to develop a broad picture of a specific city and help local public authorities in public policies with a longitudinal analysis. In 2017, Porto Alegre had 74 hotels listed in TripAdvisor, of which 60 were analyzed for having passed the 100-comment limit.

Table 1 presents some sample characteristics. While Hotel stars are given by an official classification system from MTur, TripAdvisor Stars are calculated based on reviews. About the former, the sample comprises seven hotels with two stars, 42 with three stars, and 11 with four stars, most of them being budget hotels and 19 belonging to hotel chains. Most of them are located in downtown (Centro), a region that concentrates the main touristic attractions and events in the city. Hotel names are not mentioned here.

Table 1

**Hotel sample characteristics**

Code	TripAdvisor stars	Hotel chain?	Neighborhood	Hotel stars
H1	4.5	No	Moinhos de Vento	3
H2	5	No	Floresta	3
H3	4.5	No	São João	4
H4	4.5	Yes	Bom Fim	3
H5	4	No	Moinhos de Vento	4
H6	4	No	Bela Vista	4
H7	4	No	Centro	4
H8	4	No	São João	3
H9	4	No	São João	3
H10	4	Yes	Praia de Belas	3
H11	4	No	Centro	3
H12	4	No	Centro	3
H13	4	Yes	Moinhos de Vento	4
H14	4	No	Moinhos de Vento	3
H15	4	No	Petrópolis	4
H16	4	Yes	Centro	2
H17	4	No	Centro	2
H18	4	No	Cidade Baixa	2
H19	4	No	Bela Vista	3
H20	3.5	Yes	São João	3
H21	4	Yes	Floresta	3
H22	4.5	No	Centro	2
H23	4	Yes	Navegantes	3
H24	4.5	No	Floresta	3
H25	3.5	No	Floresta	3
H26	4	Yes	Sarandi	3
H27	4	Yes	Independência	3
H28	3.5	No	Centro	3
H29	3.5	No	Centro	2
H30	4	No	Centro	4
H31	4	No	Centro	3
H32	4	No	Floresta	3
H33	4	No	Centro	4
H34	3.5	Yes	São João	3
H35	3.5	No	Centro	3
H36	3.5	No	Floresta	3

Continues

**Table 1 (continued)**

Code	TripAdvisor stars	Hotel chain?	Neighborhood	Hotel stars
H37	3.5	No	Moinhos de Vento	3
H38	3.5	Yes	Centro	3
H39	3.5	Yes	Centro	2
H40	3.5	Yes	Cidade Baixa	3
H41	3.5	No	Centro	3
H42	3	Yes	Navegantes	3
H43	3.5	Yes	São Geraldo	3
H44	3.5	No	Centro	3
H45	3	No	Cidade Baixa	3
H46	3	No	Centro	3
H47	3	No	Centro	3
H48	3	No	Petrópolis	3
H49	3	No	Petrópolis	3
H50	3	No	Centro	3
H51	3	Yes	Cidade Baixa	3
H52	3	No	Centro	3
H53	3	No	Floresta	3
H54	3	Yes	Centro	3
H55	2.5	No	Centro	2
H56	4	No	Centro	4
H57	4	No	Rio Branco	4
H58	4	Yes	Centro	4
H59	4	No	Petrópolis	3
H60	4	Yes	Cidade Baixa	3

All hotels use the TripAdvisor platform, considering that there is at least one response from the hotel managers to at least one guest comment. The replies were not collected, since the research objective was to analyze the opinion of the customer/guest. Moreover, there is no possibility for the customer to reply to the hotel manager, i.e., the customer-hotel interaction is over after the manager's response to the guest's initial comment.

Another data collected was the monthly average occupation rate (AOR) of each hotel, provided by hotel associations, in the same period. AOR data was available until April 2017.

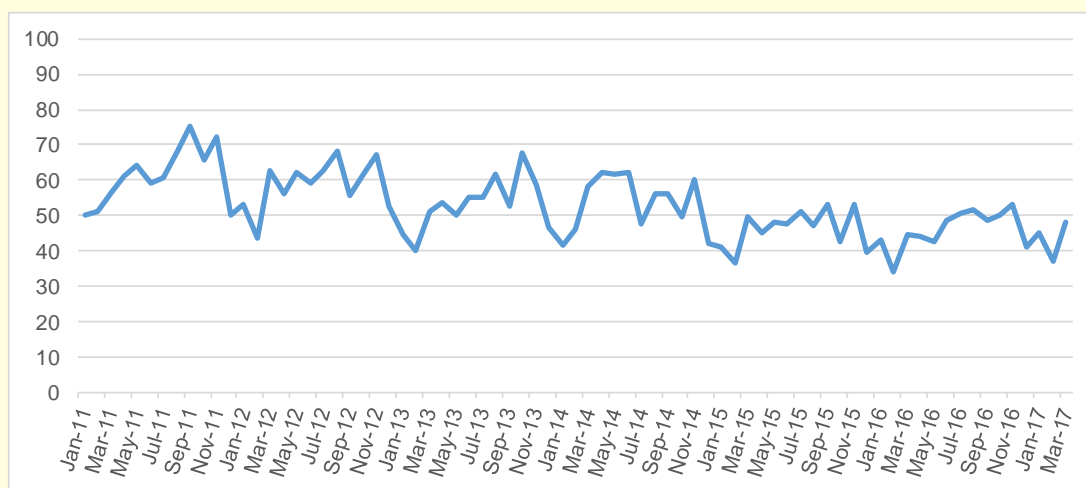
The complete textual data set contains 26,141 valid comments (considered as documents). For this study, we worked on comments from 2011 to 2016, in order to correlate with the available occupation data, resulting in 23,229 records. The average comment per hotel was 387. Table 2 presents the amount of comments per year and the mean occupation rate by year, ranging from 61.18% in 2011 to 46.01% in 2016.

Table 2

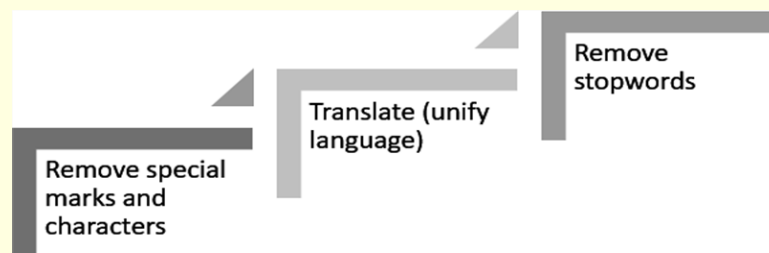
**Mean occupation rate**

Year	Mean (%)	Comments (Qty.)
2016	46.01	7,530
2015	46.22	5,991
2014	53.67	4,939
2013	53.04	3,630
2012	58.76	743
2011	61.18	396
Total		23,229

Porto Alegre is a city recognized for the quality of services. In relation to hotel services, Porto Alegre has about 250 types of hosting (like hotels, hostels and house renting), offering around 14,000 beds (Secretaria Municipal do Turismo de Porto Alegre [SMTUR-POA], 2017). Porto Alegre receive many travelers throughout the year, except during vacations and special dates (such as holidays, conferences, conventions, world forums, exhibitions, concerts and cultural festivals). Figure 2 shows the AOR per month.

**Figure 2.** Average occupation rate

Pre-processing steps followed previous works (Marcolin & Becker, 2016). This pre-processing has three important steps: removal of special marks and characters, translation (language unification), and removal of stopwords, as depicted in Figure 3.



**Figure 3.** Pre-processing steps

Source: Marcolin, C., & Becker, J. (2016, August). Exploring latent semantic analysis in a big data (base) (p 3). *Proceedings of Americas Conference on Information Systems*, San Diego, CA, USA, 22.

The first procedure was the removal of special characters and accent marks. Both procedures were performed with a script in R, which is available from the authors upon request. Examples of special characters removed are ©, \*, #, and so on. These characters may have been resulted from errors in data collection or even some language misuse from the user. Also, in this step, we removed special Portuguese marks, like á, é, ù and others, by replacing them with the clean vowel (i.e., a, e, u, and so on).

The second procedure was to unify the language of the comments. Records were found in English, Spanish, German and Portuguese. All titles were translated into English, with the translation being done using the Translate formula available in Google Sheets, which is able to handle more than 23,000 lines, albeit with some difficulty.

Finally, the third procedure of the data pre-processing step was the removal of stopwords, that is, words with high frequency in the database but without significant value. For the Term-Document Matrix construction and for other subsequent procedures, an automated script was used in R, through RStudio software.

## Results and Discussion

With more than 20,000 dimensions and a sparsity rate of more than 90%, the full Term-Document Matrix is an example of the high numbers that may result from this kind of data. In order to analyze them, the first task is to choose the parameter  $k$ , reducing the number of dimensions in the latent semantic space. An optimal  $k$  would permit to work with a fair dimension reduction, which can reduce noise in latent semantic space and retain the main dimensions that are related with the highest singular values.

This can lead to a richer relational structure that reveals latent relations presented between documents and terms (Bergamaschi & Po, 2014). But finding the optimal  $k$  is still a challenge. Different authors have proposed a number of solutions (Bergamaschi & Po, 2014; Kulkarni, Apte, & Evangelopoulos, 2014; Wild, Stahl, Stermsek, & Neumann, 2005), but many of them refer that this point should be defined empirically for each collection.

The first data exploration was related to the main topics contained inside the database, per year. For each data subset, we chose a  $k$  value through a singular-value analysis, as in Figure 4. It is

possible to see the decreasing curve of singular values, indicating that to work with all dimensions would imply more computational cost than information value. For our database, we first removed sparse terms, and after we chose to retain 65% from all singular values, as recommended in previous studies (Wild et al., 2005), ending up with a database structure as in Table 3.

Table 3

**Database**

Year	Documents	Unique Terms	Dimensions
2016	7530	354	140
2015	5991	365	143
2014	4939	337	182
2013	3630	339	181
2012	743	260	119
2011	396	342	122

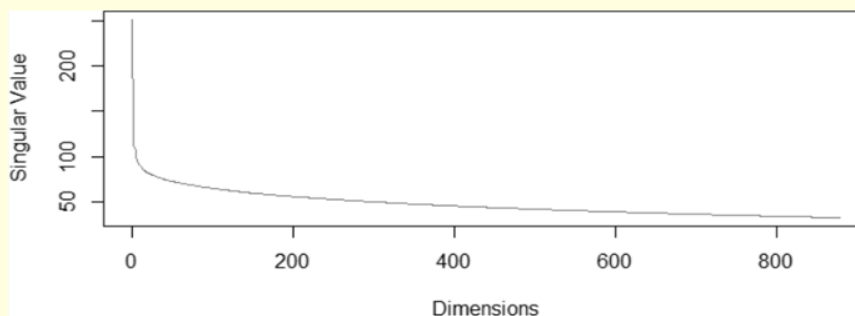
**Figure 4.** Singular values from 2011 subset

Table 4 presents the five main topics from the 2016 comments, that represents the five dimensions with the highest singular value from this subset. We can see that each of them brings a different topic to discussion. T1, that represents Topic One, has comments that reinforce the proximity from the hotels to the airport in Porto Alegre as an important message for other users. T2 brings another discussion, focusing more on attendance and reporting staff qualities like helpful and attentive. T3 comments are more concerned with hotel localization, highlighting the proximity with a restaurant or a shopping mall. Although including again the word airport, it is possible to see that T4 is different from T1, since that word is put together, in the LSA space, with shuttle and transfer, which means that comments on this dimension were more concerned with getting from and to the city airport. T5, finally, shows the cost-benefit from city center hotels, since there is a hotel pole near the main bus station.

The highlighting of these terms confirms the results presented by Xu, Wang, Li and Haghighi (2017), who noted that staff, room, location, and value, for example, are key attributes that affect customer satisfaction. At the same time, such elements are also key attributes that affect customer dissatisfaction. The validation of these aspects in different studies reveals that the decisions of the hotel managers should prioritize these attributes, as they are critical to customers experience.

Table 4

**Main topics – 2016**

T1	T2	T3	T4	T5
porto	excellent	located	airport	cost
alegre	staff	center	near	benefit
airport	great	old	shuttle	bus
center	helpful	shopping	transfer	center
near	attentive	restaurants	free	station
best	super	close	comfortable	quality
ibis	Well	bus	great	access
city	restaurant	well	good	easy
easy	wonderful	city	price	price
access	comfortable	bars	excellent	simple

This is just one way to observe the data. After the pre-processing steps, to construct a Term-Document Matrix with 99% of sparsity and an LSA space with 65% of the singular values, it took no more than a few seconds to process in a notebook with a Core i5 processor and 8 GB of memory. After that, the manager can understand, in an objectively way, the main topics that their customers (or another hotel's customer) are talking about.

Aiming to understand the changes in topics through the years, by using the same previous processes we identify the main topics present in online guest reviews in 2015. Table 5 presents the five main topics from 2015 and 2016 comments, which represent the five dimensions with the highest singular value from these subsets, as well as the changes throughout the years. The unique terms (that are presented only in one year) are highlighted in bold.

Table 5

**Main topics – 2015/2016**

2015					2016				
T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
<b>Room</b>	center	ibis	old	alegre	porto	excellent	located	<b>airport</b>	<b>cost</b>
<b>bathroom</b>	alegre	<b>standard</b>	center	porto	alegre	staff	center	<b>near</b>	<b>benefit</b>
<b>Bed</b>	porto	<b>neighborhood</b>	alegre	best	<b>airport</b>	great	old	<b>shuttle</b>	<b>bus</b>
comfortable	located	restaurants	porto	<b>one</b>	center	<b>helpful</b>	shopping	<b>transfer</b>	center
Old	restaurants	<b>network</b>	<b>bad</b>	<b>stayed</b>	<b>near</b>	<b>attentive</b>	restaurants	<b>free</b>	<b>station</b>
<b>shower</b>	shopping	shopping	<b>night</b>	excellent	best	<b>super</b>	<b>close</b>	comfortable	<b>quality</b>
<b>Large</b>	great	bars	simple	<b>stay</b>	ibis	well	<b>bus</b>	great	<b>access</b>
<b>spacious</b>	city	<b>budget</b>	city	<b>time</b>	city	restaurant	well	<b>good</b>	<b>easy</b>
Staff	bars	<b>windmills</b>	<b>historic</b>	<b>perfect</b>	<b>easy</b>	<b>wonderful</b>	city	<b>price</b>	<b>price</b>
<b>View</b>	well	<b>new</b>	located	<b>always</b>	<b>access</b>	comfortable	bars	excellent	simple

**Note.** Terms in bold represent the ones that are different from one year to another.



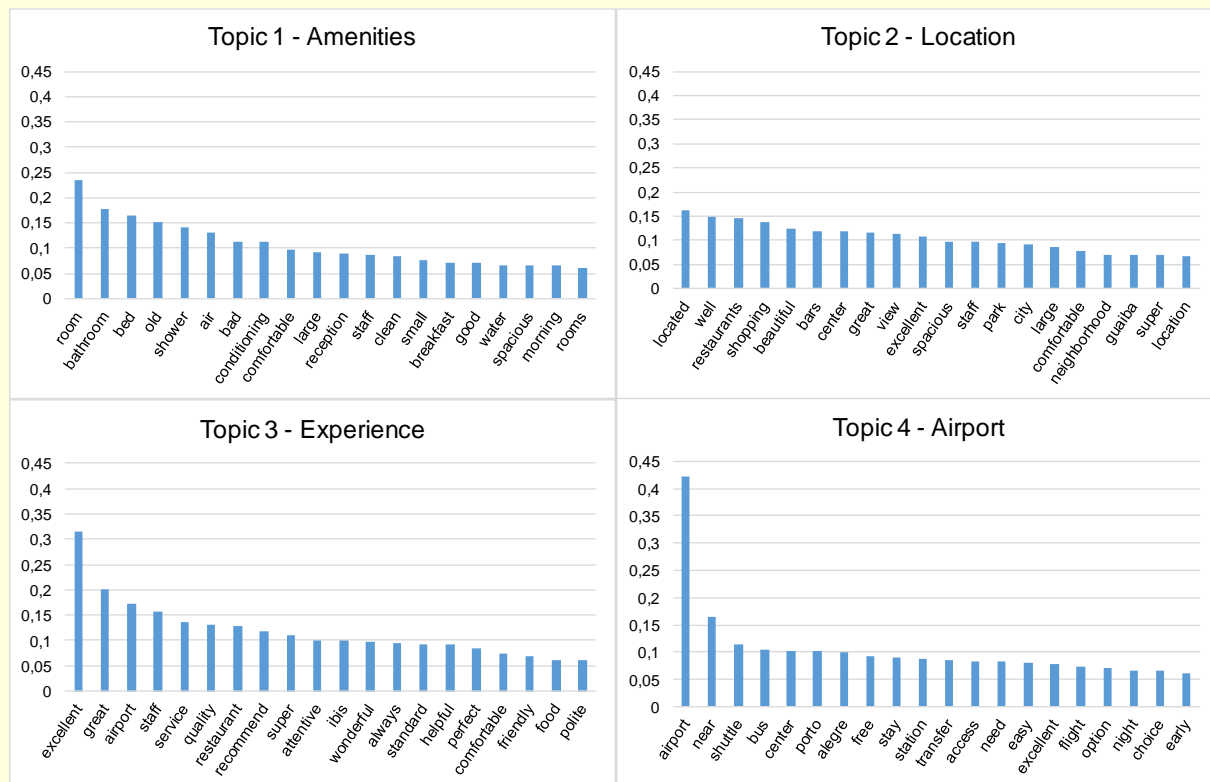
Some interesting conclusions can be drawn from this comparative table. First, it is possible to note that most 2015 topics differ from those in 2016. While topic 2, in 2015, is like topic 3, in 2016, other topics are distinct. Besides that, considering that the table shows the first five dimensions, i.e., the five eigenvectors connected with the highest singular values considering the 2015 subset, we can conclude that airport was a concern that have grown up from 2015 to 2016, representing a 50% difference between the terms from 2015 to 2016 (in bold).

In relation to the different dimensions, topic 1 in 2015, room, represents comments that have mentioned hotel room elements like bathroom, bed and shower, altogether with qualities like comfortable, large and spacious, or with complaints like old. The topic that brings the word ibis, T3, represents an expressive set of comments that mentioned this specific hotel chain. It is important to note that those are not comments screened by hotel, but by topic, and the related words - like standard and neighborhood - represent the terms that appear in the same dimension. Finally, the last topics, T4 and T5 in 2015, refer to center (downtown), with the main characteristics expressed by the users like old, simple or bad, but Porto Alegre city with qualities like best, excellent and perfect and opinions like stay(ed) and always, showing that the users who stayed in downtown Porto Alegre, despite not liking that neighborhood, liked the city as a whole.

Just like airport was not among the first dimensions in 2015, room and the related terms were not among the first dimensions in 2016. Since there was evidence that some differences throughout the years existed, we considered important to analyze the full period, aiming to understand the presence of topics in a longitudinal way. In order to do that, we needed to work with the full matrix. That was a challenge in itself, one that many organizations will have to uncover in order to increase the amount of data for decision making: the dimension reduction problem.

Our full matrix consisted of 22,062 words distributed in 23,229 documents (i.e., comments). But this does not mean that we had more than 450 millions entries to deal with. For example, approximately 60% of the words (13,205 terms) appeared only once across all years and all documents. This illustrates the semantic structure presented by Zipf law (Zipf, 1949), that states that the inverse relationship among frequency and rank-position given a frequency table for any corpus is true in different languages, which implies the presence of a similar structure for any set of documents.

We chose to remove those 13,205 terms, in order to reduce the initial dimensionality, as those words would hardly imply some global knowledge about travelers' opinion. Another procedure was to recalculate the Term-Document Matrix by removing highly sparse terms after TF-IDF computation. With that, we were able to work with a matrix 96% sparse, against the 100% sparsity that existed before, with sparsity being the number of zero-filled entries given all matrix entries. Finally, when developing the LSA space, we tested three different possibilities, with 100% (all dimensions), 50% and 65% of the singular values, and chose to stay with the latter due to costs and benefits considering data variance and computational cost (Visinescu & Evangelopoulos, 2014). After that, we ended up with 351 terms distributed in 196 topics. The main topics are presented in Figure 5.

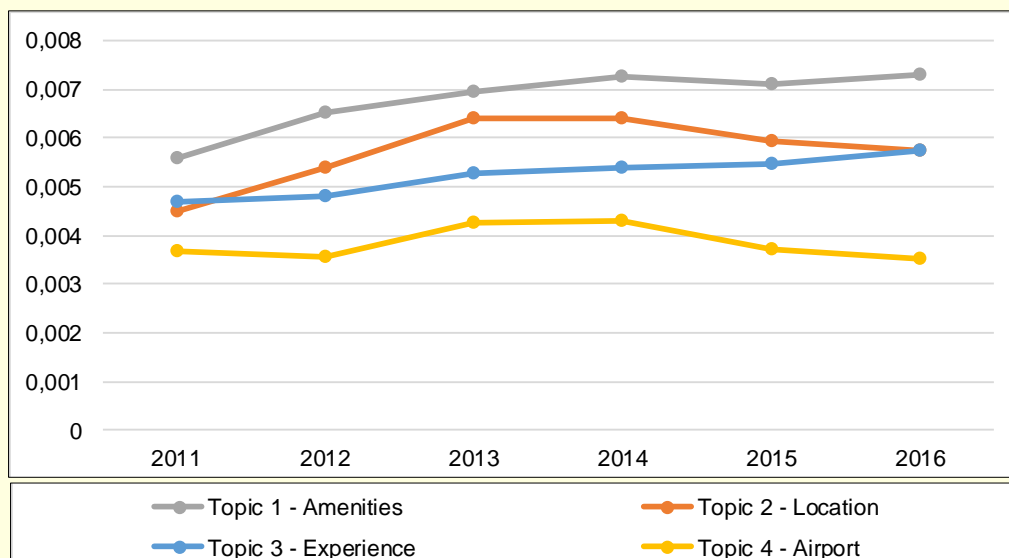


**Figure 5.** Main topics

The main topics are like the results found by Han et al. (2016). Both studies find the topics experience (terms associated with a guest's experience, including positive descriptive terms and specific characteristics, like excellent and staff, for example) and location (terms associated with the location or vicinity of a hotel, including restaurants and shopping, for example). Topic amenities in this study can also be associated with topic amenities by Han et al. (2016), although the latter has terms associated with the amenities provided by the hotel like breakfast and wi-fi (more general terms about the hotel), while amenities in our research relates to terms that are more specific to the room, like bathroom, bed, shower and air. Han et al. (2018) still highlight topic transactions (mechanics of a guest's stay) and value (guests' perceived value or money), while our research finds topic airport (related to location), differences that might be explained by cultural and geographic factors.

In this sense, the top four topics resemble findings in the study of Han et al. (2016), showing how some topics may have similar characteristics in different studies, such as experience, to refer to the customer experience, be it positive or negative, and revealing some issues that are more pleasant or unpleasant to customers. At the same time, we perceived the presence of topics different from the ones in Han et al. (2016), thus being an example of how geographic and cultural issues change from one place to another (Vieira et al., 2018). Such factors directly influence a guest's stay, and the hotel can play an important role in ensuring a better customer experience (Xu et al., 2017).

In order to represent the main topics in the full period, the top-20 words were used. We were able to understand their presence over the years, as presented in Figure 6.



**Figure 6.** Topics and years

We can see that the topics have different behaviors during the years. Although from one year to another it is hard to tell what is trending and what is out of fashion, by analyzing the topics during the years, and specially by comparing the trending curves, it is easier to retrieve information for the decision-making process.

The intense use of eWOM platforms is challenging companies in the tourism and hospitality industry, that needs to constantly respond to guests' evolving demands. The massive adoption of such platforms changes the logic of management, that becomes more dynamic (Del Vecchio et al., 2017; Horner & Swarbrooke, 2016). In order to gain competitive advantage, it is important to quit the passive logic of acting only after a complaint or a public exposure happens, thus being able to anticipate customers' concerns. For that, comparing discussion topics in a longitudinal view, as presented here, may be crucial. For example, while location and airport are declining topics in recent years, which may mean that these aspects are declining in relative importance, the room and experience topics have increasing presence, thus these aspects may be also increasing in importance for the customer's decision. In addition, amenities remains in the first place through the years, which could mean that hotel managers should give a second thought before thinking of those aspects as cost cut opportunities.

The ability to anticipate concerns also give hotel managers the opportunity to surprise their guests, which can be seen as an important aspect in the hospitality and tourism industry (Yoo et al., 2016). Being the topic trends estimated straight from reviews by an unsupervised approach, it is possible to state that a topic rise or decline can signal a safe direction for further investment aiming to increase customer satisfaction. For example, the declining of the location and airport topics might indicate the impact of the new low-cost transportation options (Bashir, Yousaf, & Verma, 2016), such as Uber and Cabify, which may explain why a hotel's location and distance

from the airport are not on top of customer interest. As such, shuttle services can be revisited in a cost-benefit approach. On the other hand, the topics room and experience are strategic factors of the hotels' internal environment, and their growth represents the importance of these factors for the customers in recent years (Perez-Aranda, Anaya-Sanchez, & Ruizalba, 2017). Also, with new work structure and cheaper tickets, people may be travelling more, what in turn may be increasing the importance of comfort issues, as travelers spend more time out of home. Thus, a safe strategy for hotel managers could be to invest in these aspects to ensure customer satisfaction.

The use of unsupervised learning combined with a longitudinal analysis allows not only for anticipative strategies, but also helps to understand how natural language can be positioned as a valuable resource. With this method, it is possible not only to overcome the challenges of natural text - that keeps many companies away from adopting text-based solutions -, but also to foster its application for strategic and market analysis, as it opens the possibility for more resources to improve decision making . In this direction, the adoption of an open-source solution that helps to capture, clean and visualize publicly available textual data from the customer can help organizations of different sizes, thus giving rise to new forms of competition based more on data analysis than data access or other available resources (Langley & Leyshon, 2017).

## Conclusions

In information technology, information has a growing value for any business-related context. The tourism and hospitality industry is an example, where there is a need for customer feedback in order to improve the quality of services. This is because guest reviews spontaneously express experiences, opinions, feelings and concerns, thus revealing relevant issues to support the decision-making process of hotel managers. As an intangible asset, information on opinions, positions, issues for improvement and other factors is important not only for customer retention, but also because of the word-of-mouth effect.

With the explosive expansion of social media, that effect grew exponentially, as millions of new customers have gained access to feedback resources through platforms like the one analyzed in our study - TripAdvisor. This paper contributes with the analysis of such data in a Business Analytics scenario, by demonstrating different forms to extract knowledge from publicly available data. Our objective was to identify the main evaluation topics presented in online guest reviews, reveal changes through the years, and uncover how to analyze an expressive amount of data. For that, we collected and treated 23,229 comments from 2011 to 2016. We used LSA in order to extract the main themes or topics from the set of comments. Besides analyzing the data, another contribution was the development of an automated script in R that, with just the hotel's URL, can scrap all comments already posted. The code and the data used here are available upon request to the first author.

In addition, the findings aim to help managers to understand the importance of analyzing large amounts of data to support their decisions. By considering raw textual data, it was possible to identify that some specific points in the whole hotel experience are more in evidence, with reviews covering amenities, location, experience and airport issues as the top aspects commented by

customers. Besides understanding the main aspects in a given region, we also explored the changes in volume of mentions throughout the years, thus revealing that some aspects may increase in incidence while others decrease. This can lead to market-oriented strategic decision making, thus helping to prioritize some operations and improve the performance of hotel attributes to meet guest demands in a faster manner. Thus, this article contributes to theory by demonstrating how to combine unsupervised learning and longitudinal analysis to make market trends evident and by using publicly available customer textual data.

A limitation of this study is the lack of individual analysis by type of hotel. For that reason, a next step would be to compare a hotel's main topics with the full industry's topics, so as to better understand the competition landscape. Future studies could also use the same tools to analyze different tourism sectors that have an online presence, like touristic points and restaurants, thus helping their managers and local governments to develop strategic actions in line with the perspective of customers.

## Acknowledgments

The authors gratefully acknowledge the sponsorship of National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Level or Education Personnel (CAPES). The authors are also thankful to Federal University of Rio Grande do Sul (PPGA/UFRGS) for supporting this research.

## References

- Acito, F., & Khatri, V. (2014). Business analytics: Why now and what next? *Business Horizons*, 57(5), 565-570. <https://doi.org/10.1016/j.bushor.2014.06.001>
- Aggarwal, C. C., & Zhai, C. X. (2012). *Mining text data*. New York, NY: Springer Science & Business Media.
- Ashton, T., Evangelopoulos, N., & Prybutok, V. (2014). Extending monitoring methods to textual data: A research agenda. *Quality & Quantity*, 48(4), 2277-2294. <https://doi.org/10.1007/s11135-013-9891-8>
- Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval* (Vol. 463, 2<sup>nd</sup> ed.). New York, NY: ACM Press.
- Bashir, M., Yousaf, A., & Verma, R. (2016). Disruptive business model innovation: How a tech firm is changing the traditional taxi service industry. *Indian Journal of Marketing*, 46(4), 49-59. <https://doi.org/10.17010/ijom/2016/v46/i4/90530>
- Bayrak, T. (2015). A review of business analytics: A business enabler or another passing fad. *Procedia - Social and Behavioral Sciences*, 195, 230-239. <https://doi.org/10.1016/j.sbspro.2015.06.354>
- Bergamaschi, S., & Po, L. (2014). Comparing LDA and LSA topic models for content-based movie recommendation systems. In V. Monfort, & K. H. Krempels (Eds.), *WEBIST: Web information systems and technologies - Lecture notes in business information processing* (Vol. 226, pp. 247-263). Cham, Switzerland: Springer.
- Cantalops, A. S., & 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>
- Carrasco, R. A., & Villar, P. (2012). A new model for linguistic summarization of heterogeneous data: An application to tourism web data sources. *Soft Computing*, 16(1), 135-151. <https://doi.org/10.1007/s00500-011-0740-1>

- Crain, S. P., Zhou, K., Yang, S. H., & Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In C. Aggarwal, & C. Zhai (Eds.), *Mining text data* (pp. 129-161). Boston, MA: Springer.
- Davenport, T. H., & Dyché, J. (2013, May). *Big data in big companies*. Retrieved from <http://datascienceassn.org/sites/default/files/Big%20Data%20in%20Big%20Companies%20-%20Tom%20Davenport.pdf>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6%3C391::AID-ASI1%3E3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9)
- Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from social big data: Implications for smart tourism destinations. *Information Processing & Management*, 54(5), 847-860. <https://doi.org/10.1016/j.ipm.2017.10.006>
- Ferreira, D. (2019). Research on big data, VGI, and the tourism and hospitality sector: Concepts, methods, and geographies. In M. Sigala, R. Rahimi, & M. Thelwall (Eds.), *Big data and innovation in tourism, travel, and hospitality* (pp. 75-85). Singapore: Springer.
- Han, H. J., Mankad, S., Gavirneni, N., & Verma, R. (2016). What guests really think of your hotel: Text analytics of online customer reviews. *Cornell Hospitality Report*, 16(2), 3-17.
- He, W., Tian, X., Tao, R., Zhang, W., Yan, G., & Akula, V. (2017). Application of social media analytics: A case of analyzing online hotel reviews. *Online Information Review*, 41(7), 921-935. <https://doi.org/10.1108/oir-07-2016-0201>
- Horner, S., & Swarbrooke, J. (2016). *Consumer behavior in tourism* (3rd ed.). New York, NY: Routledge.
- Instituto Brasileiro de Geografia e Estatística. (2008). *Economia do turismo: Uma perspectiva macroeconômica 2000-2005*. Retrieved from <https://biblioteca.ibge.gov.br/visualizacao/livros/liv37902.pdf>
- Kim, W. G., & Park, S. A. (2017). Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance? *International Journal of Contemporary Hospitality Management*, 29(2), 784-802. <https://doi.org/10.1108/ijchm-11-2015-0627>
- Kulkarni, S. S., Apte, U. M., & Evangelopoulos, N. E. (2014). The use of latent semantic analysis in operations management research. *Decision Sciences*, 45(5), 971-994. <https://doi.org/10.1111/deci.12095>
- Langley, P., & Leyshon, A. (2017). Platform capitalism: The intermediation and capitalisation of digital economic circulation. *Finance and Society*, 3(1), 11-31. <https://doi.org/10.2218/finsoc.v3i1.1936>
- Lee, M., Jeong, M., & Lee, J. (2017). Roles of negative emotions in customers perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. *International Journal of Contemporary Hospitality Management*, 29(2), 762-783. <https://doi.org/10.1108/ijchm-10-2015-0626>
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. New York, NY: Cambridge University Press.
- Marcolin, C., & Becker, J. (2016, August). Exploring latent semantic analysis in a big data (base). *Proceedings of Americas Conference on Information Systems*, San Diego, CA, USA, 22.
- Martin, D. I., & Berry, M. W. (2011). Mathematical foundation behind latent semantic analysis. In T. L. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35-55). New York, NY: Routledge.
- Molinillo, S., Sandoval, J. L. X., Morales, A. F., & Stefaniak, A. C. (2016). Hotel assessment through social media: the case of TripAdvisor. *Tourism & Management Studies*, 12(1), 15-24. <https://doi.org/10.18089/tms.2016.12102>
- Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3), 583-595. <https://doi.org/10.1016/j.ejor.2014.08.029>
- O'Leary, B. S., Lindholm, M. L., Whitford, R. A., & Freeman, S. E. (2002). Selecting the best and brightest: Leveraging human capital. *Human Resource Management*, 41(3), 325-340. <https://doi.org/10.1002/hrm.10044>

- Olmos, R., Jorge-Botana, G., Luzón, J. M., Martín-Cordero, J. I., & León, J. A. (2016). Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system. *Information Processing & Management*, 52(3), 359-373. <https://doi.org/10.1016/j.ipm.2015.12.002>
- Perez-Aranda, J., Anaya-Sanchez, R., & Ruizalba, J. (2017). Predictors of review sites usage in hotels. *Tourism & Management Studies*, 13(2), 52-59. <https://doi.org/10.18089/tms.2017.13205>
- Ransbotham, S., & Kiron, D. (2017). Analytics as a source of business innovation: The increased ability to innovate is producing a surge of benefits across industries. *MIT Sloan Management Review*, 58(3), 1-16.
- Secretaria Municipal do Turismo de Porto Alegre. (2017). BEMTUR - Boletim estatístico municipal do turismo em Porto Alegre. Retrieved from [http://lproweb.procempa.com.br/pmpa/prefpoa/turismo/usu\\_doc/bemtur\\_encarte\\_2017.pdf](http://lproweb.procempa.com.br/pmpa/prefpoa/turismo/usu_doc/bemtur_encarte_2017.pdf)
- Spangler, S., & Kreulen, J. (2007). *Mining the talk: Unlocking the business value in unstructured information*. Indianapolis, IN: IBM Press.
- 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>
- Tang, C., & Guo, L. (2015). Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (eWOM) communication. *Marketing Letters*, 26(1), 67-80. <https://doi.org/10.1007/s11002-013-9268-8>
- Thomaz, G. M., Biz, A. A., Bettoni, E. M., Mendes-Filho, L., & Buhalis, D. (2017). Content mining framework in social media: A FIFA world cup 2014 case analysis. *Information & Management*, 54(6), 786-801. <https://doi.org/10.1016/j.im.2016.11.005>
- Thorleuchter, D., & Van Den Poel, D. (2012). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, 39(17), 13026-13034. <https://doi.org/10.1016/j.eswa.2012.05.096>
- Tinkler, S., & Woods, J. (2013). The readability of principles of macroeconomics text-books. *The Journal of Economic Education*, 44(2), 178-191. <https://doi.org/10.1080/00220485.2013.770345>
- Turrel, A., Speigner, B., Djumalieva, J., Copple, D., & Thurgood, J. (2019). Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings. In K. G. Abraham, R. S. Jarmin, B. Moyer, & M. D. Shapiro (Eds.), *Big data for 21<sup>st</sup> century economic statistics* (pp. 419-454). Chicago, IL: University of Chicago Press.
- Vieira, D. P., Hoffmann, V. E., & Alberton, A. (2018). Public investment, competitiveness and development: A study into Brazilian tourism destinations. *Brazilian Journal of Public Administration*, 52(5), 899-917. <https://doi.org/10.1590/0034-7612174959>
- Visinescu, L. L., & Evangelopoulos, N. (2014). Orthogonal rotations in latent semantic analysis: An empirical study. *Decision Support Systems*, 62, 131-143. <https://doi.org/10.1016/j.dss.2014.03.010>
- Wickham, H. (2015). *Rvest package demonstration*. Retrieved from <https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/>
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. P. Lajoie, & M. Vivet (Eds.), *Artificial intelligence in education* (Vol. 99, pp. 535-542). Amsterdam, Netherlands: IOS Press.
- Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. *Proceedings of CAA International Computer-Assisted Assessment Conference*, Loughborough University, Loughborough, UK, 9.
- Xu, X. (2018). Does traveler satisfaction differ in various travel group compositions? Evidence from online reviews. *International Journal of Contemporary Hospitality Management*, 30(3), 1663-1685. <https://doi.org/10.1108/ijchm-03-2017-0171>
- Xu, X., Wang, X., Li, Y., & Haghghi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management*, 37(6), 673-683. <https://doi.org/10.1016/j.ijinfomgt.2017.06.004>
- Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1), 180-182. <https://doi.org/10.1016/j.ijhm.2008.06.011>

- Yen, C. L., & Tang, C. H. (2015). Hotel attribute performance, eWOM motivations, and media choice. *International Journal of Hospitality Management*, 46, 79-88. <https://doi.org/10.1016/j.ijhm.2015.01.003>
- Yoo, K. H., Sigala, M., & Gretzel, U. (2016). Exploring TripAdvisor. In R. Egger, I. Gula, & D. Walcher (Eds), *Open tourism* (pp. 239-255). Berlin, Germany: Springer-Verlag Berlin Heidelberg.
- Zhao, Y. (2013). *R and data mining: Examples and case studies*. Cambridge, UK: Academic Press.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Boston, MA: Addison-Wesley Press.


## Author contributions

- 1<sup>st</sup> author: data curation (equal); formal analysis (lead); funding acquisition (equal); investigation (lead); methodology (equal); software (lead); writing-original draft (lead).
- 2<sup>nd</sup> author: methodology (lead); resources (equal); validation (lead); writing-review and editing (equal).
- 3<sup>rd</sup> author: data curation (equal); resources (equal); supervision (lead); writing-review and editing (equal).
- 4<sup>th</sup> author: conceptualization (equal); formal analysis (supporting); funding acquisition (equal); visualization (equal); writing-original draft (supporting).
- 5<sup>th</sup> author: conceptualization (equal); project administration (lead); visualization (equal); writing-original draft (supporting).

## Authors


### Carla Marcolin

Universidade Federal de Uberlândia, Faculdade de Gestão e Negócios  
Av. João Naves de Ávila, 2121, 38408-100, Uberlândia, MG, Brazil  
cbmarcolin@gmail.com

 <https://orcid.org/0000-0003-0260-5073>


### João Luiz Becker

Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo  
Av. 9 de Julho, 2029, 01313-902, São Paulo, SP, Brazil  
beckerjoaoluiz@gmail.com

 <https://orcid.org/0000-0003-4176-7374>


### Fridolin Wild

Oxford Brookes University, Department of Computing and Communications Technologies  
Headington Campus, OX3 0BP, Oxford, UK  
wild@brookes.ac.uk

 <https://orcid.org/0000-0001-7344-9800>


### Giovana Schiavi

Universidade Federal do Rio Grande do Sul, Escola de Administração  
Rua Washington Luis, 855, 90010-460, Porto Alegre, RS, Brazil  
giovanaschiavi@hotmail.com

 <https://orcid.org/0000-0002-8032-5598>

### Ariel Behr

Universidade Federal do Rio Grande do Sul, Escola de Administração  
Rua Washington Luis, 855, 90010-460, Porto Alegre, RS, Brazil  
behr.ariel@gmail.com

 <https://orcid.org/0000-0002-9709-0852>

