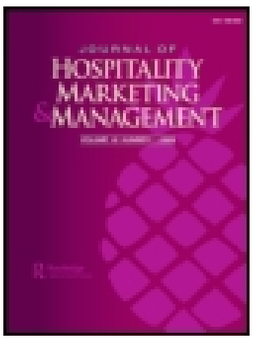

Routledge
Taylor & Francis Group

Journal of Hospitality Marketing & Management

ISSN: 1936-8623 (Print) 1936-8631 (Online) Journal homepage: http://www.tandfonline.com/loi/whmm20

Sentiment Classification of Consumer Generated Online Reviews Using Topic Modeling

Ana Catarina Calheiros, Sérgio Moro & Paulo Rita

To cite this article: Ana Catarina Calheiros, Sérgio Moro & Paulo Rita (2017): Sentiment Classification of Consumer Generated Online Reviews Using Topic Modeling, Journal of Hospitality Marketing & Management, DOI: 10.1080/19368623.2017.1310075

To link to this article: http://dx.doi.org/10.1080/19368623.2017.1310075

Sentiment classification of consumer generated online reviews using topic modeling

Ana Catarina Calheiros,[1] Sérgio Moro,[2,3,*] and Paulo Rita[4,5]

Abstract

The development of the Internet and mobile devices enabled the emergence of travel and hospitality review sites, leading to a large number of customer opinion posts. While such comments may influence future demand of the targeted hotels, they can also be used by hotel managers to improve customer experience. In this article, sentiment classification of an eco-hotel is assessed through a text mining approach using several different sources of customer reviews. The latent Dirichlet allocation modeling algorithm is applied to gather relevant topics that characterize a given hospitality issue by a sentiment. Several findings were unveiled including that hotel food generates ordinary positive sentiments, while hospitality generates both ordinary and strong positive feelings. Such results are valuable for hospitality management, validating the proposed approach.

Keywords Sentiment classification, hospitality, customer reviews, text mining, topic modeling.

[1] ISCTE Business School, ISCTE – University Institute of Lisbon, Tel: +351 21 790 30 24
[2] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal, Tel: +351 210 46 40 23
[3] ALGORITMI Research Centre, University of Minho, Guimarães, Portugal
[4] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal
[*] Corresponding author. Email: scmoro@gmail.com; Address: Av. Forças Armadas, 1649-026 Lisboa, Portugal
[5] NOVA Information Management School (NOVA IMS), Campus de Campolide, 1070-312 Lisbon, Portugal

1. INTRODUCTION

1.1. Decision support in hospitality

Hospitality traditionally lags other sectors in adopting information technology, but this has been changing in recent years and research into its multiple applications has followed suit (Šerić et al., 2014). The hospitality industry is crucial for the economy and thus it is a subject of great interest for researchers in different domains, such as management science, marketing and information technologies. Competition in this industry had an effect on client related areas, with hotels increasing investment in customer retention, customer relationship management (CRM) and targeting (e.g., Karakostas et al., 2005). The unparalleled growth of Internet applications to travel and tourism has produced a surplus of new opportunities and challenges to consumers and practitioners. Travelers generally tend to conduct an online search about their preferred destinations in order to make their travel decisions (Xie et al., 2014). Afterwards, they can read and use the reviews as references to make solid-grounded decisions about their next destination and where to stay.

For hotels to survive in today's changing business environment, hotels' managers need to have a continuous focus on solving challenging problems and exploring new opportunities. That demands an investment in computerized managerial support decision making, which implies the need of decision support and business intelligence systems (Sharda et al., 2017). The most advanced of such systems usually include the application of data mining for exploring hidden patterns in data that can be translated into useful knowledge. Data mining problems can be

addressed through machine learning supervised techniques, if there is a target to model, or unsupervised techniques, if the goal is to find relations between instances of the problem addressed (Sharda et al., 2017). An example of the former is the detection of unsatisfied customers to prevent churning through the use of decision trees (Maier & Prusty, 2015), while the categorization in segments of customers regarding their personal tastes toward hotel websites through association rules represents an example of the latter (Leung et al., 2013). Nevertheless, most of data mining approaches in hospitality are linked to forecasting tourism demand (Moro & Rita, 2016)

1.2. Text mining and sentiment classification

Text mining is a particular type of data mining that consists in analyzing textual contents for unveiling the hidden patterns that may be translated into actionable knowledge (Fan et al., 2006). The textual contents may include documents, comments, reviews or any other sort of related information, constituting the corpus for feeding text mining tasks. Typical tasks include text categorization, text clustering and sentiment analysis, among others (Srivastava & Sahami, 2009). Through sentiment analysis and classification, hospitality managers are able to better understand which characteristics among the services offered in their units can influence most satisfaction of their customers, hence helping in the definition of CRM strategies. The study by Gan et al. (2016) is an example of sentiment analysis of online restaurant reviews, as it unveiled sentiments related to food, service and context affecting more the overall given review score when compared to price. Such knowledge can be useful to restaurant managers investing their efforts in improving the former characteristics instead of focusing on adjusting prices.

The advent of Internet social media has triggered sentiment classification, a recently developed web mining technique that can perform analysis on sentiments or opinions based on published online reviews and comments (Költringer & Dickinger, 2015). It aims to extract the text from written reviews for certain products or services by classifying them into positive or negative opinions according to the polarity of the review (Cambria et al., 2013; Casaló et al., 2015). Sentiment classification can be categorized in: (1) machine-learning approaches, which rely on widely used machine learning techniques such as neural networks; and (2) lexicon-based approaches, by adopting sentiment lexicon of pre-defined sentiment related terms (Medhat et al., 2014). However, there are fewer studies based on the latter. Gao et al. (2015) tested the accuracy of three proprietary lexicon-based web-service tools applied to online reviews extracted from TripAdvisor: (1) AlchemyAPI, a software as a service API; (2) Semantria, a multilingual sentiment engine; and (3) Text2Data, a scalable API service. The same authors concluded that AlchemyAPI is the best tool to classify hotel reviews. Gascón et al. (2016) also conducted a lexicon-based sentiment analysis of the online reviews of four hotels by crossing a semantic-based dictionary with the different characteristics of the hotels (e.g., rooms, staff, location). They suggest that an evaluation of sentiment polarity (i.e., negative versus positive sentiments) can help determine the marketing strategy of the hotels; hence, such knowledge can prove to be valuable for understanding the competition's strategy. Nevertheless, most studies on sentiment classification found in the literature are focused on improving the accuracy of sentiment classification, not unveiling the products and services targeted by customers' reviews (e.g., Shi & Li, 2011). One exception is the research published by Xiang et al. (2015), which adopted a text mining approach to analyze reviews posted on Expedia.com. They used statistical methods

for measuring word frequency to infer on guests' experience and satisfaction. However, they consider single words only, not including multiple word terms, thus no n-gram analysis was conducted such as the approach suggested by Soper & Turel (2012). Also, such an immense corpus of reviews could benefit from a clustering algorithm that would gather the reviews in logical groups to provide managers with insights about users' feedbacks.

1.3. Proposed approach

Rossetti et al. (2015) followed a text mining approach to provide recommendations to specific users based on their previous online reviews. Based on words' frequency, they conducted topic modeling to make judged recommendations, also providing means to get feedback from users about the recommendations made. At the end of the aforementioned study, the authors propose using topic modeling for sentiment classification as a possible future research direction to explore. The present study aims at filling such gap in hospitality research. The latent Dirichlet allocation (LDA) allows modeling different topics, previously defined according to the distribution of the different terms across reviews (Blei, 2012). This modeling technique is the most widely topic modeling method used and it has also been used in the approach proposed by Rossetti et al. (2015). It allows determining the probability of the chosen review belonging to each topic, grouping reviews according to their proximity regarding each considered term. It also helps in identifying which topics are capturing more attention and in finding gaps for future research.

The main contributions of the present article are:

- Providing a scalable sentiment classification procedure applied to a specific hotel unit;

- Using the LDA topic modeling to discover the feelings generated by several hotel issues, thus crossing both semantics of sentiment polarity and hotel domain;

- Applying the proposed method to an eco-hotel unit to find topics that may unveil how guests' satisfaction is being perceived, hence providing valuable knowledge for hotel managers' to understand the strengths and weaknesses of the specific unit.

Next section describes the materials such as the reviews used and also the methods proposed and applied. Section 3 presents and discusses the results. Finally, the last section is devoted to the conclusions, limitations, and recommendations for further research.

2. MATERIALS AND METHODS

2.1. Data collection

For this empirical research, relevant information about a Portuguese eco-hotel from different data sources was collected, mainly the *Areias do Seixo* hotel (http://www.areiasdoseixo.com/en/hotel-overview.html). This "green" hotel was selected to illustrate that in a small hotel, with a different concept, as it is known as a "thematic luxury eco-hotel", it is also possible to apply advanced decision support systems based on text mining for supporting hoteliers managerial decisions and enhance hotel effectiveness; thus, implementation of this solution does not depend on hotel type or size. Such information comprised 401 different

review comments related with the total of 3,179 reservations during January and August of 2015, which were used as the main input for the experimental procedure. The sample size is comparable to the study by Pekar & Ou (2008) employing a sentiment analysis technique for evaluating 268 reviews of major hotels based on customer's reviews posted on the website "epinions.com", using attributes such as food, room service, facilities, and price to automatically analyze customer sentiments towards those features.

The collection of comments came from six different sources, as shown on Table 1. For all these sources the period considered was the January-August 2015 timeframe. Taking into consideration the dispersion of reviews about this hotel in both on-line and off-line domains, all the six sources were included. Such decision was also emphasized by the hotel manager's request to strengthen the study in order to present an overall picture of customers' feedback, independently of the platform used for writing the reviews. In fact, the hotel manager participated in the process of collecting the reviews to assure the reliability of the reviews included. By offering his valuable insights on the business, it was possible to unveil that sometimes the same comment was written in two distinct data sources, namely TripAdvisor and the guest's book; to address this issue, the comment was attributed to half (0.5) for each of the two sources, justifying the decimals seen in Table 1. TripAdvisor is considered one of the most popular and well-known travel and vacation services micro blog website (O'Connor, 2008) and its impact has been analyzed in previous studies (e.g., Filieri, 2015). The guest's book contributed with most of the reviews; therefore, it was included, even though it is an offline hand written source. The strategy of providing a guest's book makes customers write down their opinions when they leave the hotel and since they are not forced to do so (as it is the idea behind

a normal questionnaire), it brings more spontaneity and will to write a review. Other online sources of information from different evaluation websites besides TripAdvisor were also considered; these included some important platforms such as Zomato, a mobile application used by customers to discover the best places to eat. Apart from websites, follow up emails (emails that the hotel send to customers asking for feedback) and emails sent directly and spontaneously from customers to the hotel, elucidating hotel managers on the strengths and weaknesses, were also included.

In the beginning of this research, an unstructured interview took place with one of the hotel managers for obtaining his perceptions on the customer feedback, as well as to understand the main strategy followed by this particular eco-hotel, in order to provide in-depth knowledge for analyzing the results. This valuable in-loco knowledge allowed to strengthen the discussion of the results. For supporting this study, several different sources of customers' reviews were included in the analysis, while on one hand this may be seen as a strength by providing a more holistic view of customer feedback, on another hand it might be seen as a particular limitation of the study. Additionally, it should be emphasized that not only online reviews were used, which is the normal course of reviews-based research, but also offline sources were included as a way to enrich data, namely the guest's book. This source was included taking into consideration the dispersion of reviews about this hotel and to focus on a different approach which puts emphasis in an *in loco* feedback when analyzing customers' opinions. Also, taking into consideration that two thirds of the data came from the guest's book, it is important to highlight and address this limitation, which may translate into a higher number of positive comments from this source, considering that the guest's book is an open book where the guest is free to give his/her opinion

about the stay, meaning that most people only tend to write down positive comments. Another hint related to the different behavior regarding this source is the fact the book only contained 401 different opinions, from the total 3,179 reservations made. This means that some of the remaining 2,778 reservations during the same period could have been negative. Therefore, such limitation was addressed by including also reviews from online sources such as TripAdvisor, where customers tend to write not only positive comments but also negative ones. Also, it should be highlighted that the hotel manager conducted an evaluation of the final dataset by extracting random samples and performing a consistency analysis of those reviews. Such procedure strengthened the reliability of the data gathered.

2.2. Knowledge extraction

Figure-1 shows the approach proposed for extracting useful knowledge from the unstructured text contained within customers' reviews. This approach is based on the method proposed by Moro et al. (2015) for a literature analysis using a set of relevant articles on Business Intelligence applications to the banking industry. Usually, text mining involves two processes for building the corpus of reviews: cleaning the text from irrelevant words such as articles and adverbs; and stemming, to reduce words to a single root word (e.g., "feelings" is reduced to "feel"). However, the present analysis focused specifically on sentiment analysis and hospitality issues; hence, besides the collection of reviews, which constituted the main input from where hidden patterns of knowledge were extracted, the procedure was also fed with a lexicon that established the dictionary of relevant terms for both sentiment analysis and hospitality. Therefore, the cleaning and stemming processes used the lexicon contained in the dictionary to reduce the reviews to sets

of relevant terms. Some of the terms are constituted by more than one word, unlike the studies of Rossetti et al. (2015) and Xiang et al. (2015). By considering n-grams (Soper & Turel, 2012), the procedure can embed some context through the combination of a few words (e.g., "social networks").

Accordingly, two distinct dictionaries were built. Table 2 shows the dictionary for hospitality, including all the reduced terms and several of the similar terms (the remaining ones were omitted to save space). The set of base terms was extracted from the dictionary published by Ingram (2003). Then, it followed the approach from Lau et al. (2005), from where important hotel attributes were also extracted to enrich the hospitality dictionary. Finally, an analysis of a sample of the 401 reviews allowed for the identification of additional terms, which were also included. In Table 3, the sentiment classification dictionary is shown (also only a subset of the reduced equivalent terms is displayed). To define this dictionary, first an accurate classifier (scale) was required to compile indicators of sentiment; then the sentiment was determined by comparing comments against the expert-defined entry in the dictionary, which makes it easier to determine the polarity of a specific set of words. The scale used in order to develop the sentiment analysis dictionary followed the approach of Hu et al. (2012). Thereafter, the dictionary was enriched with terms representing sentiment intensifiers, following different polarity, including "strong positive";" ordinary positive"; "ordinary negative"; and "strong negative" categories.

Considering that the definition of a dictionary and grouping terms under a unique reduced term is subjective, in order to reduce such subjectivity, an independent Marketing and Hospitality specialist validated both dictionaries. The lexicon constituted from both dictionaries was

compiled as a single input, to allow a cross-domain relationship analysis between sentiments and hospitality. The main output from the text mining procedure is the document term matrix, as shown in Figure 1. This matrix has two dimensions: the reviews (usually text mining is performed over documents, hence the name) and each of the terms considered; each of the cells contains the frequency each term occurs in each of the reviews. For analysis, two user friendly outputs were provided: a table of frequencies, to count the number of occurrences of each term, and a word cloud, to provide an easier visual interpretation of those occurrences.

The document term matrix is the only input for the LDA topic modeling. The LDA is a three-level hierarchical Bayesian modeling process that groups collections of items in topics defined by identified words or terms and the probability that each of them characterizes the topic (Blei, 2012). Such model enables to analyze the relative relevance of each term using the β distribution value, which characterizes the relation between the topic and the given term. All β value are negative, thus to facilitate the interpretation, the absolute value for all cases are considered. A β closer to zero represents a stronger relation between a term and its corresponding topic. The LDA final output is a tridimensional matrix encompassing terms, reviews and topics. Therefore, for every topic it is possible to obtain a measure of its relationship to one of the dictionary terms through the β distribution. Also, for every review it is possible to check to which topic it suits better. The product of these three dimensions results in a very large structure. Considering the goal is to analyze the relation between sentiments and hospitality, only the most relevant sentiment and the most relevant hospitality issue were scrutinized.

The R statistical tool was adopted for all experiments. This is an open source tool and provides flexibility through the installation of packages published by a large number of supporters in the CRAN (Comprehensive R Archive Network - https://cran.r-project.org/). For the text mining procedures, the "tm" package was adopted, considering it was specifically developed to conduct the text mining functions needed to analyze text (Meyer et al., 2008). This package provides functions to convert unstructured into structured data, reducing dimensionality of data while keeping relevant information, and jointly analyzing quantitative and qualitative data. To gather the topics which group comments, the "topicmodels" was adopted, since it receives as input the data structures produced by the "tm" package in order to provide basic infrastructure to fit topic models (Hornik & Grün, 2011).

3. RESULTS AND DISCUSSION

3.1. Text mining

In this section, the results obtained for the text mining procedure are shown. Table 4 exhibits the number of occurrences for each of the reduced terms according to the equivalences determined from the dictionaries (Table 2 and Table 3). The grayed rows represent terms related to the four types of sentiments used to classify customers' satisfaction. Findings show that positive sentiments are ten times more frequent than their respective negative counterparts. Moreover, the text mining procedure recognized a perceived strong positive sentiment as the most frequent term of both sentiment and hospitality domains, occurring 739 times in the whole 401 reviews, with an ordinary positive sentiment being the second most frequent term, with 601 occurrences.

These results are aligned with the hotel manager's perceptions, who stated in the previous discussion that the majority of customer's verbally expressed that they were definitely happy and delightful with their stay, with the reviews confirming such claim.

Figure-2 shows the word cloud for terms from the hospitality domain only, providing a visual interpretation of the results. The sentiments were excluded considering only four sentiment classifications are considered, thus allowing a clearer picture of the relevance of hospitality terms. Thus, the cloud is drawn on the frequency of terms occurring in the textual contents of the reviews: a term occurring more frequently will be visualized in a larger font. First, it should be stressed that the term "hospitality" is accounting for accommodation related terms only, as shown in Table 2. The global results, presented in both Table 4 and Figure 2, with a total of nineteen hospitality terms, show that the words "food" and "hospitality" are clearly the main hotel attributes mentioned by customers, particularly in measuring the main reasons for customer satisfaction in this eco-hotel as a tourism destination. There is also a relevant interest in location, romance and people.

3.2. Topic modeling

The number of topics is a required parameter for computing the LDA model which can be tuned for optimal results (Yi & Allan, 2009). Following the approach of Moro et al. (2015), this value was initially set to half of the terms considered, hence twelve topics were modeled. The resulting topics included several overlapping topics if the two most relevant terms were considered, suggesting the ideal number of topics summarizing the reviews should be less than twelve. As in any knowledge discovery project, a cyclic procedure must be carried out with further iterations

until the model is tuned for achieving the best possible results (Sharda et al., 2017). Therefore, the procedure included iterative experiments by reducing the number of topics and evaluating the obtained models through the entropy measure (in an approach similar to Hornik & Grün, 2011), with the results being consistent with the reduction in the number of topics, reaching to the final tuned number of nine (Table 5) which represent the best distribution of the dataset of reviews.

Each topic is shown in horizontal lines, with the column labeled "Hospitality term" presenting the most relevant hotel attributes and the column labeled "Sentiment term" the most relevant sentiment regarding each topic, and also a column for the β distribution values in respect to a given topic (where a smaller value represents a stronger relation). The number of comments column (#) presents the number of reviews that were included in each topic. Figure 3 shows the same information from Table 5 but through a visual picture of the topics.

The most noticeable characteristic of all the topics unveiled is that all of them are related to positive feelings, with four of them representing an "ordinary positive" sentiment, and the five remaining ones representing a "strong positive" sentiment. This is a confirmation of the text mining results achieved in Section 3.1. While such result has the limitation of not showing the poorer aspects of this hotel unit for improvement, it unveils that the hotel strategy is paying off, generating positive feelings that provide feedback for other customers, in a valuable word-of-mouth communication that can potentially bring more customers.

By looking at Table 5, it is remarkable the fact that in general there is not such a high level of difference between the β values for hospitality and sentiment terms within each topic, by comparison with the results of Moro et al. (2015). The largest difference happens for the first

topic (0.53-3.01) while the remaining topics show consistent β values where the lower β values is above half the β value when compared to the other domain lower term. By comparison, Moro et al. (2015) showed results with consistently larger differences between the most relevant and the second most relevant terms (the largest being 0.03-4.35). Such result reduces the problem identified in their study about weak correlation terms, strengthening the relations discovered between the two distinct domains.

The first topic, being the most mentioned service regarding hospitality terms, is best identified with "food" and gets a matching of 65 comments, having a significant lower β value (0.53) meaning that the relation between the topic and the hospitality attribute is strong. However, the sentiment term associated presents a higher value (3.01), resulting in a distant relation from the topic. Despite the given β value for "ordinary positive", it is still the most common sentiment concerning those 65 customer reviews, regarding food. Topic number five also underlines the sentiment "ordinary positive" and its relation with "food", getting a closer β value in this topic, 1.92 on what regards "food" and 1.14 for the associated sentiment, reinforcing this relationship, enclosing a total of 104 reviews from the universe of 401 analyzed, including the first topic.

Both the first and fifth topics represent interesting discoveries, and where one can hypothesize that by targeting customers with attractive food in the service offered may also serve the purpose of retaining them by upgrading ordinary to strong positive feelings. These are expected results, considering first that food is highly related with satisfaction (Lin & Mattila, 2010), and second that the food industry employs a large number of workers, posing the difficult challenge of managing them in order to fully satisfy customers (Kruja et al., 2016). The second topic, which is

best identified with “location”, gets 58 matching comments, having a lower β value (2.24), when compared to the associated sentiment term, with a 1.24 β value, considering the “location” of this hotel as a “strong positive” sentiment expressed by customers. This puts emphasis on the number of positive adjectives expressed by customers associated with its location, within just five minutes’ walk to the beach, far enough from largely populated cities, but reachable in a 40 minutes’ drive from Lisbon, where the main Portuguese airport is located, as well as where many of the Portuguese national tourists live (http://www.areiasdoseixo.com/hotel-directions.html). Topic number six also emphasizes this “strong positive” sentiment regarding “location”, presenting very close β values for both terms, meaning a tighter relation between both terms and the topic. These are expected results, according to the study by Kandampully & Suhartanto (2000), focusing on hotel’s location as being one dimension of hotel image attributes, and as a consequence one of the most important factors considering customer intention to repurchase, recommend and exhibit loyalty.

Another important relation can be observed in the third topic, associated with “different”, expressing an “ordinary positive” sentiment, and where one can notice the strength of this relation, highlighted in Table 5. The hotel is considered to be a “thematic luxury eco-hotel”, focusing its strategy on differentiation, and customers seem to be pleased with that hospitality attribute. Nevertheless, the closer relation of both terms implies that hotel manager’s may still have room to improve and transform this ordinary into strong positive feelings. The forth topic is “romance”, gathering 45 comments, and also presenting a strong relation between both the terms and the topic. It shows a β value of 2.07 for romance, and a closer value for the expressed sentiment term (1.11); such a result translates that customers express a “strong positive”

sentiment regarding romantic characteristics and details of the place. This represents a good investment from the hotel manager's considering the hotel's web site presents several offers and vouchers related to romance (http://www.areiasdoseixo.com/hotel-products.html).

Topic number seven stresses "hospitality" and its relation with an "ordinary positive" sentiment. This is a strong relation, looking at its approximate β values, 1.91 for "hospitality" and 1.44 for the combined sentiment. However, topic number nine, even though taking into consideration the same "hospitality" term, presents a different sentiment (with 37 different customer opinions) where the sentiment emphasized in this case corresponds to "strong positive", highlighted by its lower β values, meaning that this is a strong relationship. This relation with the topic "hospitality" implies that customers value the quality of the hotel's amenities and its services and it is one of the customer satisfaction factors in this hotel, but it still leaves room for further improvements.

It should also be stressed the interesting result displayed by topic number eight, with the hospitality term "site" getting a match of 35 comments, where customers express a "strong positive" sentiment. According to the hotel manager during the initial interview, one of the main reasons for people to choose this hotel was to see a picture of the place in the Internet or in an international magazine, read an opinion of an influencer travel blogger or on TripAdvisor and also searched for *Areias do Seixo* website. Although being one of the main reasons for people to choose this place, it is also one of the main hospitality products contributing for a higher customer satisfaction in this hotel, according to the results shown in Table 5. One consideration can be made from the nine topics. Location, hospitality and food are the terms which best

identify with six of them, totalizing 269 comments from the total of 401 (around 67%), making of location, hospitality and food the main valued hospitality terms mentioned by customers.

3.3. Discussion of theoretical and practical implications

Despite results focusing on several different topics that can be characterized by a specific sentiment, as seen in Table 5, the same results conceal certain limitations. One of them consists of the fact that the given results do not show an emphasis on "people" nor on "decoration", which is an unexpected result since it is considered by the hotel manager as one of the main reasons for people to visit the place. Furthermore, it is considered to be a Green Hotel, making great efforts in environmental issues and initiatives. However, in the results such environmental awareness is not mentioned as a top issue in any of the topics found; this is an unexpected result, as the hotel is considered a well-known and respected eco-hotel. Moreover, this is considered a key differentiation strategy of the unit, although the term "different" appears highly related to "ordinary positive", showing there is still plenty of room for improvement regarding their core strategy. According to Richard Hammond, founder of Green Traveller, "sustainability is still not a top criteria for choosing a holiday destination, things such as location, price and facilities are still the main drivers; however, being green has established a secondary consideration adding value to final customer, and making this a growing market tendency" (http://eandt.theiet.org/magazine/2011/07/eco-hotels.cfm).

One relevant implication of the experiments conducted in this study is the fact that the results achieved using LDA topic model are presenting the same "hospitality term" twice for some topics. Food, hospitality and location are the examples of such issue, which may be justified

considering the automated nature of the process. Nevertheless, it also reveals that some hospitality related terms are emphasized in detriment of others. Another important implication that should be discussed consists of the fact that no negative sentiment characterized topic was found, limiting the value to enhance the satisfaction perceived by customers, even though negative reviews are a less representative subset, according to the topics found. Furthermore, considering that the sample represents only a single hotel unit in Portugal, the specific hotel issues identified in customer reviews obviously reflects the perceptions of location-related aspects of this hotel. Sentiment classification and guest satisfaction would probably be considerably different in another cultural context. Another limitation relates to clustering algorithms, as some reviews could eventually be associated to two distinct topics. This issue was addressed by minutely analyzing the most relevant comment for each topic with the purpose of confirming the given results, as explained in Section 3.4. Nonetheless, these potential limitations do not reduce the internal validity of data and thus do not harm the purpose of demonstrating the power of sentiment mining techniques in the field of hospitality.

In general the processed results state that customers appreciate the place, for its unique location, the quality of its amenities, romantic surroundings and characteristics of the place, as for the hotel's positive reviews and Internet visibility. However, it is notable that the services offered and attributes such as food, hotel amenities and difference as a hotel still have room for improvements, as management recommendations. These results are valuable for hospitality management, supporting decision making to improve the value perceived by customers, validating the proposed approach.

This study highlighted how hotel reviews may be harnessed for supporting hotel managers, whether by identifying issues caring for improvement, or by validating the efforts of marketing strategies carried out for promoting the unit. An example of the former is the identification of room for improvement regarding food, while the latter is expressed within two dimensions unveiled through the topics found: (1) the romance investment of the hotel is fruitful, being clearly recognized by guests; (2) the core strategy as an eco-hotel needs further attention, as the manager was expecting the differentiation factor to be much more emphasized by customers ("different" has been found to be highly related to "ordinary positive").

Using reviews from customers willing to express their comments without any demand (contrary to surveys) has proven to be a useful means for obtaining reliable feedback. However, reading all the reviews can be a challenging task. Text analysis techniques are valuable tools for helping to extract knowledge from a large set of reviews. While most applications rely on pure machine learning techniques, using a lexicon-based approach for sentiment classification has been lesser studied. Furthermore, topic modeling is able to link a sentiment lexicon with relevant categories within hospitality for unveiling topics of interest that may help to devise new marketing strategies as well as identify the strengths and weaknesses of on-going strategies. This is a confirmation of the usefulness of the research direction identified but not followed by Rossetti et al. (2015).

3.4. Reviews analysis

In Section 3.2, topic modeling allowed to identify interesting insights on customers' perceptions and satisfaction under the disclosed topics, characterized by the respective terms, as shown in

Table 5. However, automated approaches as the one proposed conceal a few limitations, such as the fact that topic modeling is completely dependent on the technique used for creating the topics, which is based on term identification; hence, some terms may have different meanings based on the remaining text, where subjectivity detection can be a challenging task (Thomas et al., 2011). In this section, such issue is addressed by identifying and analyzing one representative comment for each topic, as show in Table 6 (only part of the comment is displayed, due to space constraints).

The approach followed was based on full text manual analysis of the chosen nine reviews, in order to confirm the hypotheses suggested by the topics found, in a similar procedure of the one proposed by Moro et al. (2015). The numbering of topics is the same as for Table 5, while the column "frequency" stands for the number of occurrences of both the hospitality and sentiment terms, considering also the dictionaries expressed in Tables 2 and 3. In order to select the most relevant reviews per topic two metrics were considered: the number of different terms mentioned in each review, and the total number of times each of the sentiment term and hospitality term occurred. As an example, topic three groups a total of 52 reviews; for each of those reviews, the number of occurrences for the respective sentiment and hospitality term were extracted from the document term matrix, where the one with the higher frequency of both terms was selected. In this case, the selected review presents a frequency of 12, with 3 "different" terms, and 9 "ordinary positive" terms. This reinforces the relation between the different aspects of the entire hotel as an "ordinary positive" sentiment according to customers.

Topic one best matches with the example above, illustrated in Table 6, where "food" is the most relevant term, showed by the number of food terms mentioned (23), in contrast with only 9 regarding the sentiment term. This explains the difference in β values, as mentioned before. This review confirms the importance that food has on customer satisfaction, and therefore on the fact that sentiment classification has a positive attribute in hospitality. By looking at topic eight, one can notice that it clearly shows a weakness of the topic modeling approach: even though it groups 35 different reviews, the relation between the sentiment and the hospitality terms is quite distant. The review used as example (being the review having the larger number of occurrences regarding both terms) shows an unpleasant situation regarding "site" hospitality term. When analyzing the entire comment it couriers a "strong positive" sentiment due to overall hotel characteristics. Nevertheless, when converging specifically on "site" terms it is difficult to associate the presence of this strong sentiment regarding the corresponding topic. The client considers the existence of an extremely slow wi-fi as a negative sentiment, as being a "stressful situation" and "totally unnecessary". This topic reflects one of the major limitations of similar automated approaches based on clustering methods. According to Kumar & Sahoo (2014), "clustering aims at finding a subset of items which are more similar than others using similarity measures". However, clustering algorithms, such as the case of topic modeling, try to make approximations based on the information they convey. The ambiguities associated with language semantics pose a serious challenge for an algorithm to decide on what is the best matching topic.

4. CONCLUSIONS

4.1. Contributions

In this paper, sentiments polarity through text mining techniques of more than 400 customers' reviews were applied, as well as the identification of inherent relationships, using the latent Dirichlet allocation modeling, between two domains of variables in the hotel industry: sentiment classification and hospitality issues from a chosen eco-hotel. Figure 4 summarizes the methods followed and the main findings resulting from this study.

The uniqueness of this study relies on using unstructured data from several sources to understand customer perceptions and feelings of a single hotel on a scale that was not available through traditional guest survey studies. Also, the present research is justified by the study of Rossetti et al. (2015) considering their recommendations for future studies focusing on the usage of topic modeling to analyze sentiment classification. As so, the present study fills such void in hospitality research. Hence, it contributes to the literature in several ways. Firstly, it provides a scalable sentiment analysis process applied to a specific hotel unit, which is a fundamental contribution to Marketing strategy, namely Customer Relationship Management, becoming a fundamental process in order for hoteliers to increase competitive advantage and create intelligent customer databases. The usage of the latent Dirichlet allocation topic modeling to discover costumer feelings generated by several hotel issues when crossing both semantics of sentiment polarity and hotel domain is a major contribution of this study, considering that the novel trends and generalized opinions unveiled may be used in order to improve hospitality

business. Another contribution lies on the proposed method being applied to an eco-hotel unit where topics found may expose how guests' satisfaction is being perceived, hence providing valuable knowledge for hotel managers to understand the strengths and weaknesses of a specific unit.

From a practical point of view, this study stresses that core sentiments expressed through the mainstream hotel issues are deeply strong positive and ordinary positive. Customer retention seems to be associated with targeting, justifying customer satisfaction by the correlation between the unexpected location, the quality of its amenities, romantic characteristics of the place as for the hotel's positive reviews and Internet visibility. As such, the contribution of this study lies in providing a solid background support beyond a simple manual analysis of customers' reviews or a traditional guests' survey, thus strengthening managerial decisions to further improve the unit. As an example, management recommendations arose such as the room for improving food services, as well as the hotel's amenities and its differentiation focus strategy. This should be taken into account considering that the hotel managers had a different idea on what regards improvements, previously more focused on the environmental issues and the sustainability of the hotel, as qualities that people value nowadays. As such, some of the findings of the present study may lead to a shift in the hotel's current investment in "green technologies" toward the real perceived assets of the unit by customers, with a particular emphasis on its location (95 of strong positive reviews related to such issue) and romance (45 strong positive reviews). These results are valuable for hospitality management, validating the proposed approach.

4.2. Limitations

Nevertheless, the present study comprises several limitations and the findings should be interpreted with caution. First, no negative sentiment characterized topic was found, limiting the value to enhance the satisfaction perceived by customers, even though negative reviews are a less representative subset, according to topics found. Furthermore, considering that the sample represents only a single hotel unit in Portugal, the specific hotel issues identified in customer reviews obviously reflects the perceptions of location-related aspects of this hotel. Sentiment classification and guest satisfaction could be considerably different in another cultural context. Another limitation relies on clustering algorithms, as some reviews cannot match any topic; this issue was addressed by scrutinizing the most relevant review for each topic with the purpose of confirming the given results. Nonetheless, these potential limitations do not reduce the internal validity of data and thus do not harm the purpose of demonstrating the power of sentiment mining techniques in the field of hospitality.

4.3. Future research

Future research may consider applying a fully automated system approach, as this proposal is a hybrid method containing the efforts of computer programs and manual labor. The ideal option should aggregate both in a single system as a technological development. Overall, taking into consideration the actual globalization phenomenon and the development of new technological systems applied to management, this research presents a practical approach for the development of an innovative methodology that can conduct many companies through a remarkable marketing

strategy, characterized by customer focus and competitive intelligence. As such, this study may be seen as an example for the development of business intelligence systems applied to hospitality marketing and management.

Acknowledgements

The authors would kindly like to thank the *Areias do Seixo* manager Gonçalo Alves for his valuable cooperation on this research, and for allowing access to information on this respected and highly esteemed Portuguese eco-hotel.

REFERENCES

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. doi:10.1145/2133806.2133826

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, (2), 15-21. doi:10.1109/MIS.2013.30

Casaló, L. V., Flavián, C., Guinalíu, M., & Ekinci, Y. (2015). Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers. Journal of Business Research, 68(9), 1829-1835. doi: 10.1016/j.jbusres.2015.01.010

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9), 76-82. doi:10.1145/1151030.1151032

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. Journal of Business Research, 68(6), 1261-1270. doi:10.1016/j.jbusres.2014.11.006

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2016). A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews. Journal of Quality Assurance in Hospitality & Tourism, 1-28. doi:10.1080/1528008X.2016.1250243

Gao, S., Hao, J., & Fu, Y. (2015). The application and comparison of web services for sentiment analysis in tourism. In 2015 12th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-6). IEEE.

Gascón, J., Bernal, P., Román, E., González, M., Giménez, G., Aragón, Ó., ... & Crespo, J. (2016). Sentiment analysis as a qualitative methodology to analyze social media: study case of tourism. In Qualitative Research (CIAIQ2016), 2016 1st International Symposium on (Vol. 5, pp. 22-31).

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision Support Systems, 52(3), 674-684. doi:10.1016/j.dss.2011.11.002

Ingram, H. (2003) Dictionary of Travel, Tourism & Hospitality (3rd ed.), International Journal of Contemporary Hospitality Management, 15(7), 413-414. doi:10.1108/09596110310496079

Kandampully, J., & Suhartanto, D. (2000). Customer loyalty in the hotel industry: the role of customer satisfaction and image. International Journal of Contemporary Hospitality Management, 12(6), 346-351. doi:10.1108/09596110010342559

Karakostas, B., Kardaras, D., & Papathanassiou, E. (2005). The state of CRM adoption by the financial services in the UK: an empirical investigation. Information & Management, 42(6), 853-863. doi:10.1016/j.im.2004.08.006

Költringer, C., & Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. Journal of Business Research, 68(9), 1836-1843. doi:10.1016/j.jbusres.2015.01.011

Kruja, D., Ha, H., Drishti, E., & Oelfke, T. (2016). Empowerment in the Hospitality Industry in the United States. Journal of Hospitality Marketing & Management, 25(1), 25-48. doi:10.1080/19368623.2015.976696

Kumar, Y., & Sahoo, G. (2014). A charged system search approach for data clustering. Progress in Artificial Intelligence, 2(2-3), 153-166. doi:10.1007/s13748-014-0049-2

Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. Cornell Hotel and Restaurant Administration Quarterly, 46(3), 344-362. doi:10.1177/0010880405275966

Leung, R., Rong, J., Li, G., & Law, R. (2013). Personality differences and hotel web design study using targeted positive and negative association rule mining. Journal of Hospitality Marketing & Management, 22(7), 701-727. doi: 10.1080/19368623.2013.723995

Lin, I. Y., & Mattila, A. S. (2010). Restaurant servicescape, service encounter, and perceived congruency on customers' emotions and satisfaction. Journal of Hospitality Marketing & Management, 19(8), 819-841. doi:10.1080/19368623.2010.514547

Maier, T. A., & Prusty, S. (2015). Managing Customer Retention in Private Clubs Using Churn Analysis: Some Empirical Findings. Journal of Hospitality Marketing & Management, 1-23. doi:10.1080/19368623.2016.1113904

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. doi:10.1016/j.asej.2014.04.011

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. Journal of Statistical Software, 25(5), 1-54. doi:10.18637/jss.v025.i05

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42(3), 1314-1324. doi:10.1016/j.eswa.2014.09.024

Moro, S., & Rita, P. (2016). Forecasting tomorrow's tourist. Worldwide Hospitality and Tourism Themes, 8(6), 643-653. doi:10.1108/WHATT-09-2016-0046

O'Connor, P. (2008). User-generated content and travel: A case study on Tripadvisor. com. Information and communication technologies in tourism 2008, 47-58. doi:10.1007/978-3-211-77280-5_5

Pekar, V., & Ou, S. (2008). Discovery of subjective evaluations of product features in hotel reviews. Journal of Vacation Marketing, 14(2), 145-155. doi:10.1177/1356766707087522

Rossetti, M., Stella, F., Cao, L., & Zanker, M. (2015). Analysing User Reviews in Tourism with Topic Models. In Information and Communication Technologies in Tourism 2015 (pp. 47-58). Springer International Publishing.

Šerić, M., Gil-Saura, I., & Ruiz-Molina, M. E. (2014). How can integrated marketing communications and advanced technology influence the creation of customer-based brand equity? Evidence from the hospitality industry. International Journal of Hospitality Management, 39, 144-156. doi:10.1016/j.ijhm.2014.02.008

Sharda, R., Delen, D., & T. Efraim, (2017). Business Intelligence, Analytics and Data Science: A Managerial Perspective, 4th Edition. Pearson.

Shi, H. X., & Li, X. J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. In Machine Learning and Cybernetics (ICMLC), 2011 International Conference on (Vol. 3, pp. 950-954). IEEE. doi:10.1109/ICMLC.2011.6016866

Srivastava, A. N., & Sahami, M. (2009). Text mining: Classification, clustering, and applications. CRC Press.

Soper, D. S., & Turel, O. (2012). An n-gram analysis of Communications 2000--2010. Communications of the ACM, 55(5), 81-87. doi:10.1145/2160718.2160737

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. Research Synthesis Methods, 2(1), 1-14. doi:10.1002/jrsm.27

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. doi:10.1016/j.ijhm.2014.10.013

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. International Journal of Hospitality Management, 43, 1-12. doi:10.1016/j.ijhm.2014.07.007

Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In Advances in Information Retrieval (pp. 29-41). Springer Berlin Heidelberg. doi:10.1007/978-3-642-00958-7_6

Table 1 – Sources for the analyzed reviews

Nr.	Source	Nr. Comments	%
1	TripAdvisor	52.5	13%
2	Guest's book	272.5	68%
3	Follow up emails	40	10%
4	Evaluation website	4	1%
5	Direct emails	26	6%
6	Other	6	1%
	Total	**401**	**100%**

Table 2 – Dictionary for hospitality

Reduced term	Similar terms or from the same domain *
tourism	travel, tour, trip
hospitality	accommodation, hotel, resort, lodge
decoration	decorative, interior design, architecture
environment	nature, sustainability
holiday	vacation
food	restaurant, taste, flavors, wine, cuisine, meal
people	employees, staff, workers
location	place, sight, scenery
guests	clients, hosts, costumers
site	website, browsing, internet, social networks

relax	relaxed, calm, quiet, chill
feelings	sense, sensations
eur	euros, money, expensive, cost, price
reserve	reservation, booking, availability
friendly	kindness, caring, attentive, empathy, sympathy, pleasant
different	creativity, unique, singular, innovator, original
romance	love, romantic, passion
equipment	amenities, facilities
trends	theme, chic, exotic, hippie, style, lounge

* All terms are in lower case and separated by commas

Table 3 – Dictionary for sentiment classification

Reduced term	Similar terms or from the same domain *
Strong Positive	
	brilliant, excellent ,fantastic, phenomenal, wonderful, superb, beautiful, spectacular, delightful, memorable, remarkably, stunning
Ordinary Positive	
	cool, good, fashionable, helpful, peaceful, beauty, quality, warm, respect, tasty, recommend, spacious, pleasure, elegant, sincere, liked
Ordinary Negative	
	bad, nervous, loss, aversion, sad, difficulty, quite small, little scattered, expensive, shame, unbalanced, spoiled, apology
Strong Negative	
	terrible, awful, stupid, horrible, unfortunately, ridiculous, really hard, too long, weaknesses, very bad

* All terms are in lower case and separated by commas

Table 4 – Term frequency

#	Term	Frequency
1.	strong positive	739
2.	ordinary positive	601
3.	food	445
4.	hospitality	424
5.	location	290
6.	romance	230
7.	people	150
8.	different	133
9.	decoration	110

10.	relax	107
11.	holiday	89
12.	ordinary negative	68
13.	equipment	56
14.	environment	53
15.	feelings	51
16.	strong negative	46
17.	site	43
18.	eur	32
19.	trends	29
20.	friendly	28

21.	tourism	28
22.	reserve	26
23.	guests	9

Table 5 – Topics discovered

Topics	#	Hospitality term	B	Sentiment term	β
1.	65	food	0.53	ordinary positive	3.01
2.	58	location	2.24	strong positive	1.24
3.	52	different	1.84	ordinary positive	1.32
4.	45	romance	2.07	strong positive	1.11
5.	39	food	1.92	ordinary positive	1.14
6.	39	location	1.50	strong positive	1.57
7.	37	hospitality	1.91	ordinary positive	1.44
8.	35	site	2.42	strong positive	1.66
9.	31	hospitality	2.06	strong positive	1.18

Table 6 – Representative reviews per topic

	Review	Sentiment		Source
1	Breakfast: The choices were wonderful and it was like being in someone's kitchen helping ourselves to food. Dinner: We had tasting menu. All dishes were so creative and tasty. … We can not wait to go back to the hotel again and we do hope the quality of things we mention above will be kept well. We thank you so much for looking after us so well and we are grateful for the sweetest memories you gave us. Look forward to seeing you again.	☺	3 2	3
2	This was the perfect place to spend our honeymoon. Feels like paradise. Peaceful, beautiful, the staff is amazing. A 5 star experience! Thank you!!! Thank you	😃	6	2

3	Food very original and genuine, healthy and tasteful. I recommend the spa for the quality of its treatments. … A magnify place to be in a romantic environment, with details full of charm. It was a magical stay, very special. To come back one day.		10	2;3
4	Dear two Joanna's :-), Sergio, Carina, Maria and Philip . How are you? We had a safe trip back home but we miss you and your lovely place very much. … We will definitely come back soon and perhaps even for our marriage;-) We saw the beautiful scenery you made for the American couple and that really left us speechless. Lovely greetings from us and please keep in touch!		8	5
5	Thank you for such a great week for me and my group at Yoga Retreat. We love everything, the dunes and the bonfire whit Philip and Lucas. The dinner at the fish restaurant with Martha. Love that she stay with us. We love the warm girl in the spa Martha, and all kindness from Julia and all staff.		4	2

6	We stayed here for a short stay after a trip to Lisbon. In the future we'll certainly stay longer. The location and grounds were perfect, the rooms even better and finally the staff still beating both of those. The restaurant opened us to new foods and ways to present them. The drinks were fantastic and the beach was beautiful. The walk from the hotel to the ocean was through dunes with colorful vegetation ending with a tremendous view of the waves. We hope to make it a regular place to stay in the future.		1 0	1
7	Dear Daniella. … We appreciate the perfect reception upon our arrival and also your kindly and lovely smile! After travelling the world round finally we discovered that heaven exist (but only) on earth and in Portugal haha. The hotel was really " unhavre de paix " and a paradise. Be sure that we will inform our colleagues specialized in " Vip Travel " in Belgium. Once again 1000 x thanks.		8	5

8	We had a fantastic stay at Areias do Seixo. You have a very wonderful hotel with a great environment, delicious food and amazing rooms. The whole nature environment is also a very calm and beautiful experience. The only thing that definitely needs to be improved is the extremely slow Wi-Fi. Some website never loads and that became a very stressful situation and I think totally unnecessary. Today you expect it to just work fine. The small Navio restaurant even had a better connection and I understood when I was able to use one of the computers in the reception that you also had another faster connection.		1 1	3
9	We have seen and experienced many hotels but what the Areias do Seixo offers surpasses everything. … The team is open, friendly, warm, knowledgeable, funny, erected, restrained and makes the hotel what it is. A DREAM! THANK YOU FOR THE DREAM Ok, it is rather expensive and the selection rather modest but the hotel itself makes up all this.		1 5	1

Figure 1 – Proposed approach

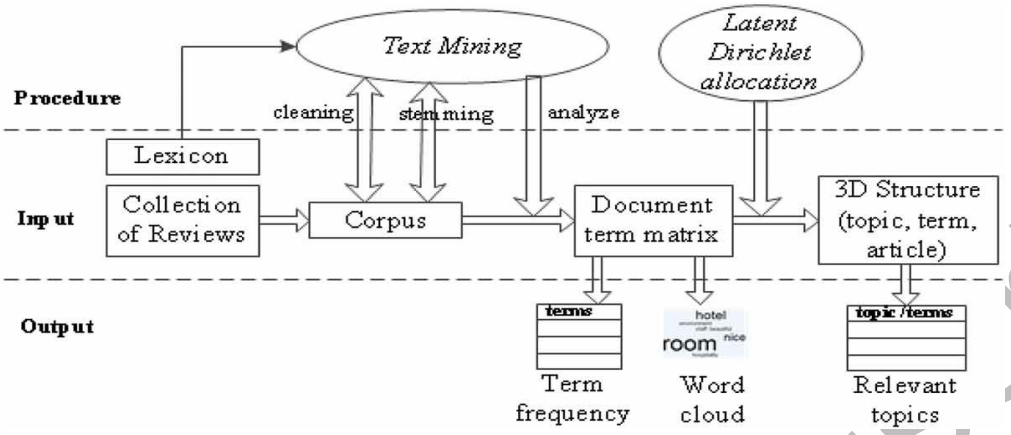

Figure 2 – Word cloud for hospitality domain

Figure 3 – Topics in a picture

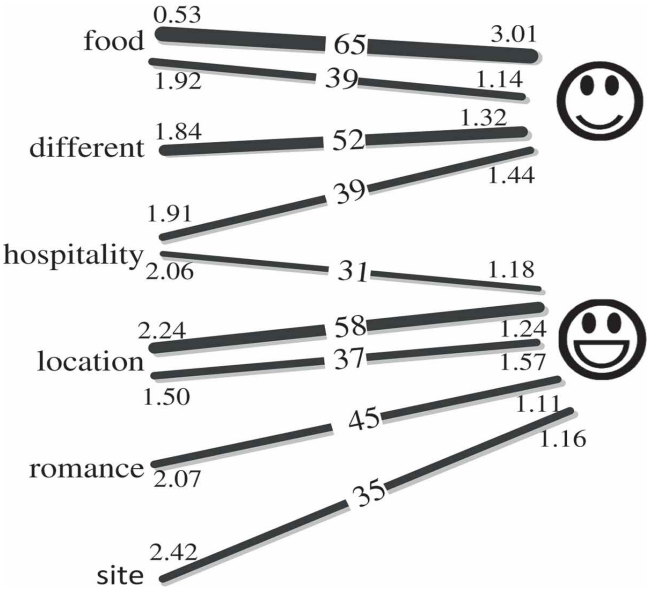

Figure 4 – Graphical summary of the methods and findings