DOI: https://doi.org/10.1016/j.ipm.2021.102645

Convolutional Neural Encoding of Online Reviews for the Identification of Travel Group Type Topics on TripAdvisor

Francisco Jose Arenas-Márquez, Rocio Martinez-Torres *, Sergio Toral

Universidad de Sevilla, Sevilla Spain

* Corresponding author: e-mail address: rmtorres@us.es (R. Martinez-Torres)

Abstract: Previous studies have concluded that there are significant differences in travelers' preferences depending on the trip type. The problem of extracting users' preferences from a corpus of text can be solved by using traditional clustering algorithms, which work quite well when there is no predefined data structure. However, in this paper, we consider the problem of extracting users' preferences when they belong to a finite number of classes represented by the trip type. In this paper, we propose an encoding method based on a Convolutional Neural Networks (CNNs), trained as a classifier for the classes that predefine data structure. The intuition behind convolutional neural encoding is its ability to maximize the distance between documents belonging to different classes in the new, derived feature space. Findings reveal that CNNs encoding has better discriminative properties than alternative encoding methods such as Latent Dirichlet Allocation or average word2vec encoding. Moreover, we demonstrate that CNNs encoding can be used to identify the unique topics associated with the predefined data structure determined, in this case, by the four trip types.

Keywords: electronic word-of-mouth; convolutional neural encoding; unique topics; customer preferences

1. Introduction

The hospitality industry has been especially impacted by the proliferation of user-generated content and electronic word-of-mouth (eWOM). Potential customers are willing to learn about other customers' experiences and judgments, as these constitute an important and integral part of the overall travel experience. Hence, they like to look through previous opinions before they make their own decisions (Mauri & Minazzi, 2013; Khorsand et al., 2020). Many previous studies have proven that information shared through platforms such as TripAdvisor and Booking play an important role in influencing travelers' decisions (Gao et al., 2017; Navío-Marco et al., 2018). However, this information can also be accessed by hotel managers, eWOM managers, and practitioners, so it can be monitored and processed to study several issues related to the tourism industry.

We can distinguish several groups of studies focused on the way that managers can benefit from shared information. The most immediate way consists of collecting and analyzing the content of reviews. Machine learning algorithms are generally used as the number of collected reviews can be enormous, so manual data analysis is not cost-effective (Xu, 2020). Recent studies have analyzed aspects such as customer preferences (Luo et al., 2020; He et al., 2020), customer dissatisfaction and complaints (Hu et al., 2019b), and even deceptive reviews (Martinez-Torres et al., 2019) using content analysis techniques such as topic modeling. A variation combining sentiment analysis and clustering methods has also been proposed for summarizing textual content to identify the top-k most useful sentences (Hu et al., 2017) and the unique attributes of tourist destinations (Toral et al., 2018). The second group of studies refers to the prediction of several issues such as customer satisfaction (Zhao et al., 2019), review helpfulness (Ma et al., 2018), hotel ratings (Gao et al., 2018), and user reputation using historical data as well as centrality and other social network analysis metrics (Martinez-Torres et al., 2018; Kumar Behera et al., 2019; Kumari et al., 2020). These studies make use of classifiers with textual reviews, customer involvement, and reviewer historical data as input features. Finally, the third group is focused on providing customers with tools to help their decision-making. For example, Peng et al. (2018) suggest an appropriate hotel decision support model to help tourists to find the best hotel, and Nilashi et al. (2018) develop a recommendation method using multi-criteria ratings.

The focus of this study falls within the first group, that is, collecting and analyzing shared information on eWOM websites. Several variants of the well-known topic modeling algorithms have been proposed to determine customers' preferences and complaints (Ding et al., 2020). However, the variety of customer profiles and types of reviews make it difficult to extract relevant topics. For example, several studies have demonstrated that review topics clearly differ depending on their sentiment orientation (Toral et al., 2018; Zhao et al., 2019; Martinez-Torres et al., 2019). One possible solution to this problem consists of analyzing positive and negative review topics separately. The problem with this approach is that some of the extracted topics can be the same for positive and negative reviews, so it would be difficult to identify whether the topics in question represent a source of satisfaction or dissatisfaction for customers. Moreover, the capability of unsupervised machine learning techniques (such as topic modeling) to extract topics from negative reviews has also been questioned, due to their topical diversity and the smaller number of negative reviews compared to positive reviews. The consequence is a lower separability of topics (Kirilenko et al., 2021). This problem becomes even more difficult when the number of classes is greater than two when separating customers according to hotel rating or trip types, for example. In this context, hotel managers would benefit from knowing the distinctive topics of interest depending on travelers' profiles. If they know this information in advance, they are able to orientate their business toward a specific traveler profile. By distinctive topics, we mean topics that are highlighted by one specific type of customer and no others. For example, business travelers may appreciate some specific services more than family travelers or couples. Consequently, hotel managers would be able to promote these services if they want to attract business travelers.

The primary aim of this paper is to propose a new methodology to extract the unique topics associated with each specific class of customer using a convolutional encoding of documents. The advantage of the proposed method is that the encoding is obtained by applying a neural classifier, so the document encoding highlights the differences between documents in different classes. The proposed methodology is applied to obtain the unique topics associated with four different trip types.

The main novelties of this paper are:

- To prove the ability of convolutional neural encoding of documents to discriminate between documents that belongs to different and predetermined classes.
- To provide a methodology able to extract those topics that can be uniquely associated to each class.
- To reveal the unique topics of interest associated to four different trip types using the content of online shared reviews. Although reviews can be separated according to the trip type, many of them address similar issues that apply across all types of hotel guests.

The remainder of the paper is structured as follows: The related work section describes how the variable trip type has been used in the previous literature as well as the different methodologies for document encoding. Section 3 formulates the research framework by proposing one research question and two hypotheses. Section 4 presents the case study and the proposed methodology, and section 5 details the results as well as the answer to the formulated research question. Section 6 discusses the findings and presents the contributions, including the managerial implications, limitations, and future works. Finally, section 7 concludes the paper.

2. **Related work**

2.1 Trip type

Considering globalization and the scenario of the travel and tourism sector, market segmentation using different trip types is becoming important for management (Middleton, 1994; Morrison, 1996; Kotler, Bowen, & Makens, 1998). Trip type means the purpose or reason for taking a trip and can be used to make recommendations (Klenosky & Gitelson, 1998). As travelers with a particular trip purpose are likely to be different than those with other trip purposes, it is generally agreed that the needs and wants of business and pleasure/personal travelers are somewhat different and result in different travel expenditure patterns (Sung et al., 2001). So, tourist behavior can be linked to leisure engagement-related theories embedded in consumer behavior research to establish marketing strategies and practices. As one of the major focuses in travel and tourism marketing research should be consumer behavior, analyzing trip-related features should offer us more comprehensible and meaningful results when interpreting people's travel behavior (Sung et al., 2001).

Many previous works have paid attention to trip types. Dimanche and Havitz (1994), in their examination of conceptual and measurement issues in the tourism sector, analyzed four areas of consumer behavior: (a) ego involvement, (b) loyalty and commitment, (c) family decision-making, and (d) novelty seeking. Hsieh, O'Leary, and Morrison (1992) divided the tourism market into the business travel market and the pleasure/personal travel market for an analysis of the Hong Kong pleasure travel market activity segmentation base. Klenosky & Gitelson (1998) studied the impact of trip type and origin on agents' destination recommendations using the following types of trips: (a) fall foliage trips, (b) historical/cultural trips, (c) honeymoon trips, (d) outdoor recreation trips, and (e) weekend getaway trips. Chu & Choi (2000) examined the extent to which six hotel selections factors were perceived as meaningful and used by business and leisure travelers. Sung et al. (2001) identified five trip types, one related to business purposes and four to pleasure/personal purposes (visiting friends or relatives, recreation, any other, and day trips) and analyzed the effect of income, travel expenditure, and demographic, sociocultural, and trip-related characteristics on the pleasure/personal trip types. Liu et al. (2013) examined the effect of trip modes, referred to as trip types by other authors, on customer expectations by analyzing customers' personal profile information and the reviews of the hotels that they visited. They found that when a traveler engages in different trip modes, there are differences in both their expectations and satisfaction. Schuckert et al. (2019) used the number of reviews, review ratings, and customer trip types to analyze the gaps between international and domestic hotels in China from the perspective of management-response strategy. Hu et al. (2019b) used the trip type variable from TripAdvisor as a control variable to investigate whether consumers highlighted different characteristics of the accommodation for different visit times and whether the factors that define customer satisfaction influence customer re-patronage in the same way. They found that customers traveling alone and/or for business purposes are more likely to re-visit a hotel, while those traveling for leisure with close companions, such as couples, families, and groups of friends, are less likely to re-visit a hotel. Yadav & Roychoudhury (2019) realized that people sharing hotel reviews tend to talk about aspects that are important to them, so they analyzed textual content using sentiment analysis to understand customers' expectations and preferences according to the trip mode. Khorsand et al. (2020) sought to obtain a process to predict future reviewer ratings for a specific hotel using trip type from TripAdvisor.com as part of the user profile information. Finally, Shen et al. (2020) referred to travel motivations for taking a trip when researching gamified trips, categorizing the reasons as seeking and escaping (Iso-Ahola, 1982); novelty seeking, self-esteem, ego enhancement, socialization, and rest and relaxation (Jang et al., 2009); and novelty and knowledge, prestigious and luxury experience, self-development, exciting experience, and escape and relationship (Li & Cai, 2012).

To sum up, trip types can be grouped into two categories: business and pleasure. The former refers to work motivations or purposes, while the latter is related to free time. This last category can be divided into different subcategories using TripAdvisor, as many authors have done: solo, couples, families, and friends (Rhee & Yang, 2015; Radojevic et al., 2017; Hu et al., 2019a; Khorsand et al., 2020).

2.2 Document encoding

The large number of textual reviews available on the Internet has triggered the proliferation of Automatic Text Summarization (ATS) systems, which are widely used as tools to obtain shortened versions of large text documents or to identify the main topics of interest within a corpus of texts (Alami et al., 2019). In natural language processing, the manipulation of documents requires their transformation into numerical vectors that can be mathematically processed. The most usual approach to transforming documents into vectors consists of the Bag-of-Words (BOW) model. In this model, documents are represented by a matrix of vectors in which each row represents the documents and each column corresponds to a word from the vocabulary of the corpus (Larcker & Zakolyukina, 2012). Documents are given by a numerical vector whose values rely on metrics based on word frequency, such as Term-Frequency (TF) and Term-Frequency Inverse-Document Frequency (TF-IDF) (Teso et al., 2018). Despite its simplicity, the BOW model has two major disadvantages. First, it results in a high-dimensional and extremely sparse representation of documents, which negatively impacts the performance of the classifiers. Second, the semantic relationships between words and the sequentiality of the text are ignored, as the words are considered independently from their position in the document (Li et al., 2016).

Some of these drawbacks can be overcome by using a generative model, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA follows a Bayesian paradigm and once it has been trained, provides the final outcomes of topic–word and document–topic distributions through a posterior maximization with Gibbs sampling. Although LDA was originally developed for topic modeling, document–topic distribution can be considered as another document representation in which both word frequencies and semantic information (topic constitution) are considered (Kim et al., 2019).

Nevertheless, the best method to preserve the semantic relationships between two words is the neural-network-based word representation method called Word2vec (Mikolov et al., 2013). This can be implemented with two different models, namely the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. The words around a target word, that is, the context, are used in both models: in CBOW, the context is the model input to predict the output, which is the target word, while in Skip-gram it is the opposite: The target word is the model input to predict the context. Document encoding can be achieved by averaging Word2Vec representation, which results in a dense representation of documents (Djaballah et al., 2019; deLira et al., 2019). An alternative is the use of neural networks to process the sequence of the Word2Vec representation of words in a document (Meškelė & Frasincar, 2020).

Document encoding is a prior step to document clustering, with the purpose of automatically organizing documents into clusters so that documents belonging to the same cluster are more similar to each other than to documents in other clusters. Document clustering involves the use of prior document encoding, which varies from sparse to dense representations (Curiskis et al., 2020). By using the similarity of vectors as the distance metric of a clustering algorithm (i.e., k-means), the final clusters can be extracted, with each cluster representing a differentiated topic in a corpus of documents. Therefore, clustering provides the final interpretation of document encoding. A clustering algorithm works quite well when the number of clusters is unknown a priori and the algorithm must identify the latent structure of data. However, in some cases, there is a predefined structure of data, so the documents already belong to a predefined number of classes (Fournier-Viger et al., 2014). This is the case of this research, where collected reviews belong to a predefined set of classes established by the trip type: family, friends, business, and couples. One possible solution consists of separately applying a clustering algorithm to each class to enable the identification of the topics related to each trip type. However, if the document encoding is not able to discriminate between classes, the obtained topics can be rather similar across several classes and impair the interpretation of the specific discussion topics associated with each class. The methodology proposed in this paper overcomes this issue by proposing a document encoding fitted during classification training.

2.3 Neural Network approaches to text classification

The emergence of neural networks is revolutionizing the field of natural language processing due to their ability to capture complex patterns beyond word similarities and temporal dependencies (Fathi & Shoja, 2018). CNNs and Recurrent Neural Networks (RNNs) have been successfully applied to make the captured patterns invariant to local translations.

CNNs are neural networks that can learn local features from words and phrases using text represented as a sequence of Word2Vec vectors. Although they were first used in a 2D version for images (height and width), they also can be used in a 1D version for sequences of words (Meškelė & Frasincar, 2020). 1D CNNs apply convolutional layers of different filter lengths to capture local features that are invariant to local translations and max-pooling layers to capture the maximum information from the text (Agarwal et al., 2018; Kim et al., 2020). The length of the filters determines the n-gram discriminating features that the network can learn. By stacking multiple convolutional layers, a higher-level encoding of the input document is obtained with the last convolutional layer providing the highest-level semantic representation of the input document (Mitra & Jenamani, 2021). The representation that this last convolutional layer provides for different input documents can be used as a semantic discrepancy measure at the document level.

The main advantage of RNNs is that they capture long-term dependencies better. Long Short-Term Memory (LSTM) networks are a specific case of RNN that are suitable for discovering such dependencies in sequences of words. This is particularly interesting for sentiment analysis applications where the presence of negative or positive terms in a completely different part of the text can change its global meaning. The main disadvantage of RNNs is that they are more prone to being affected by the vanishing gradient problem during network training (Behera et al., 2021).

Hybrid approaches using a combination of CNNs and RNNs are also popular. CNNs capture local patterns invariant to translation and RNNs can identify the temporal dependencies of such patterns. Both have been used for sentiment analysis (Behera et al., 2021) and for scoring the helpfulness of online reviews (Mitra & Jenamani, 2021). Other hybrid approaches have considered a mix of knowledge-based and deep learning techniques to perform emotion recognition and polarity detection (Cambria, 2016).

Attention mechanisms and ensemble learning have been widely used to improve model performance and increase the accuracy of predictive models. These can be optionally added at the end of previous architectures although they only provide a small improvement in accuracy (Ahmed et al., 2020). Attention mechanisms consist of focusing on the important parts of the context by assigning different weights (Basiri et al., 2021) and have been used for multi-label text classification and emotion detection problems (Ren et al., 2021; Wand et al., 2021). Ensemble learning consists of combining the predictions of multiple classifiers to reduce the risk of misclassification (Liang & Yi, 2021). There are two ways to ensemble classifiers: One is bagging, which consists of using a base classifier trained with different data subsets obtained by applying bootstrap sampling. The second is boosting, which consists of an iterative process where the weights of classifiers and the weights of misclassified data are adjusted depending on the performance of the classifiers. Generally, ensemble learning provides a better generalization capability (Cambria, 2016).

3. Research framework

Previous works have demonstrated that the trip type influences tourists' behavior and preferences. In general, customer preferences can be identified using topic modeling approaches, which basically consist of mathematically encoding documents and then applying a clustering algorithm (Martínez-Torres, M. R., & Diaz-Fernandez, 2014; Vayansky & Kumar, 2020). Topic modeling assumes that the data has no prior structure (Curiskis et al., 2020). However, in this case, there is a predefined structure in the data given by the trip type. Therefore, the problem can be formulated as the identification of topics in each class and can be solved by separately analyzing the classes. Nevertheless, the challenge with this problem is the identification of the unique topics associated with each class; for example, which topics are specifically related to business trip types and not to any other trip types? Some cross-topics exist that can be addressed independently of the trip type, such as breakfast (Moro et al., 2020). Moreover, eWOM sites like TripAdvisor require the reviewers to rate a variety of hotel features, such as location or cleanliness, so reviewers are likely to mention these topics in their shared reviews (Tsai et al., 2020). Consequently, the identification of unique topics requires eliminating cross-topics and retaining those that are specifically associated with each class. These specific topics represent the most relevant information on customer preferences. Understanding tourist preferences and behavior is crucial for hotel managers to develop a marketing strategy that meets tourists' expectations (Vu et al., 2020). Hence, we propose the following research question:

RQ1: Is there any technique that allows the identification of the unique topics associated with each trip type?

Finding the unique topics associated with each class relies on the selection of an appropriate document encoding. Several encoding methods have been reported in the literature, the most relevant of which is the Word2Vec approach. However, none considers a prior distribution of documents in several classes. This is the point where the ability of neural networks to find relevant features associated with classification problems can be exploited. Text classification problems make use of several 1D convolutional layers through which the neural network learns its own features (Ji et al., 2020). These convolutional features can be used as document encoding, with the property of having been obtained as part of a classification problem. As such, mathematical encoding emphasizes the differences between classes. Hence, we hypothesize:

H1: The convolutional neural encoding of documents provides the encoding that best discriminates between classes

There is a broad range of criteria that tourists might consider when selecting and booking hotels (Nie et al., 2020). The four main criteria that affect tourists' hotel selection are location, price, facilities, and cleanliness (Lockyer, 2005), and these are precisely the four attributes that users can rate separately on TripAdvisor. Merlo and de Souza Joao (2011) added other elements to this list such as room size, building type, service quality, silence in rooms, air-conditioning, and living environment. More recent studies have also included tourist experiences and perceptions such as tangible and sensorial experiences, staff performance, esthetic perception (Ren et al., 2016), room quality, staff attitude and behavior, access, and food (Xu & Li, 2016). However, all the previous attributes can differ depending on who tourists are traveling with (e.g., friends, family, or as a couple), or why they are traveling (whether for leisure or business, etc.) (Liu et al., 2013; Yadav & Roychoudhury, 2019), that is, they are perceived differently depending on the trip purpose. For example, Chu and Choi (2000) realized that business travelers are more concerned about the room and the front desk, while for leisure travelers security is the greatest consideration; Rajaguru & Hassanli (2018) found significant differences between leisure and business guests' perceptions of value for money and service quality at hotels with different star ratings, as did Hu et al. (2019b). The expectation of experiences is also affected by the purpose of the trip (Walls et al., 2011). For example, travelers on a business trip could have high expectations for service, while they could be less sensitive to service levels when traveling independently (Liu et al., 2013). Previous studies mainly distinguish between leisure and business trip types. However, we hypothesize that differences can be extended to more specific categories such as business, family, friends, and couples, as also considered by TripAdvisor. Content-based review classifiers provide evidence that there are some content differences able to predict the trip type (Lahlou et al., 2013). Hence, we hypothesize:

H2: Each trip type can be characterized by its own specific topics different than other trip types

4. Case study and methodology

The dataset is formed of opinions submitted to TripAdvisor about hotels located in Barcelona belonging to 4 different classes according to the trip type review: traveled with family (1430), traveled with friends (1401), traveled as a couple (1426) and traveled on business (1405). The size of the dataset is determined by the following considerations: (1) we have considered hotels located in Barcelona with more than 200 reviews written in English. The reason for this minimum value is to guarantee a variety of reviews belonging to different trip types and also to avoid hotels with a low number of reviews where fake or deceptive posts can negatively impact the results of the study, and (2) the number of reviews per trip type should be balanced as our proposed methodology relies on classifiers that work better with balanced datasets. Therefore, reviews belonging to classes with higher values were randomly discarded to ensure that the final dataset was balanced.

Figure 1 shows word distribution per document. Most of the documents contain between 30 and 200 words. Data were collected using a web scraper and accessing the fields “title of the review” and the “review body”. These were then concatenated into a single field.

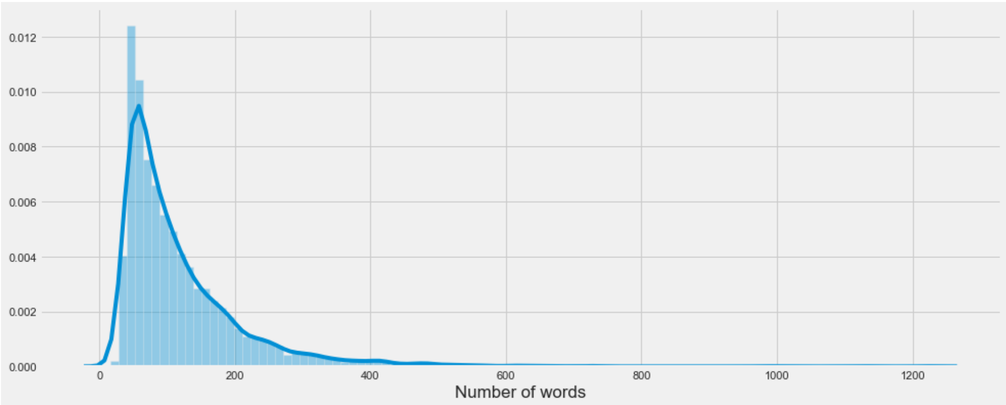


Figure 1. Distribution of the number of words per document.

Collected reviews were first pre-processed. As is usual with text mining applications, documents were converted to lower-case and then punctuation and stopwords (words that do not carry any information such as prepositions, pronouns, and articles) were removed. Text pre-processing was performed using the NLTK (Natural Language Toolkit) library in Python, which is the leading platform for building Python programs to work with human language data. Next, the text was tokenized and stemmed using the same library. Tokenization means splitting the text into their word units and stemming refers to transforming words into their roots by removing derivational affixes. In this paper, we use the traditional Porter Stemmer, which is the standard stemmer used in NLP and Information Retrieval tasks (Porter, 1980). Table 1 shows the transformation of the original reviews into their stemmed versions with several examples. The final aim of all these pre-processing stages is to transform the original texts into homogeneous versions so that every occurrence of a word is only counted once in its root form, irrespective of its linguistic form. A total of 13301 words were obtained as a result of the proposed pre-processing. The number of words is then refined by removing very infrequent words. More specifically, words occurring fewer than 10 times were removed as they were of no use for travel type identification. The final number of words was, therefore, 1956.

Review class	Original review	Stemmed review
Traveled with friends	Good Hotel, great location. Overall a very nice hotel to stay in. The pool bar was unattended but you were able to get drinks on the 1st floor & bring to the roof top pool area which is fab. Overall staff are excellent, rooms comfortable and any request was looked after. Located in a nice central area, staff are knowledgeable and friendly. Very good B/fast.	“good”, “hotel”, “great”, “locat”, “overal”, “nice”, “hotel”, “stay”, “pool”, “bar”, “unattend”, “abl”, “get”, “drink”, “1st”, “floor”, “bring”, “roof”, “top”, “pool”, “area”, “fab”, “overal”, “staff”, “excel”, “room”, “comfort”, “request”, “look”, “locat”, “nice”, “central”, “area”, “staff”, “knowledg”, “friendli”, “good”, “bfast”
Traveled as a couple	An excellent hotel in the heart of Barcelona. A charming hotel some 50 metres from La Rambla. The Guelle Palace is just across the street. With a beautiful roof terrace with a view. Charming settings, helpful staff and spacious rooms. A stunning replica of Gaudi's mosaics in the lobby.	“excel”, “hotel”, “heart”, “barcelona”, “charm”, “hotel”, “50”, “metr”, “rambla”, “guell”, “palac”, “across”, “street”, “beauti”, “roof”, “terrac”, “view”, “charm”, “set”, “help”, “staff”, “spaciou”, “room”, “stun”, “replica”, “gaudi”, “mosaic”, “lobbi”, “gaudi”, “fan”, “choos”, “room”, “face”, “street”, “palac”, “guell”, “unforget”
Traveled with family	Convenient, clean, and more. My daughter and I stayed here Friday June 28th after a	“conveni”, “clean”, “daughter”, “stay”, “friday”, “june”, “28”,

Review class	Original review	Stemmed review
	cruise. We arrived at 9am expecting to have to wait until afternoon to get into our room. Not so! Our room was ready and was a pleasant surprise. The room was well air conditioned, large and nicely decorated. The hotel is close to hop on bus line as well as Los Rambles? The only complaint I had was that the lighting in the bathroom is very poor for putting on make up.	“th”, “cruis”, “arriv”, “9”, “expect”, “wait”, “afternoon”, “get”, “room”, “room”, “readi”, “pleasant”, “surpris”, “room”, “well”, “air”, “condit”, “larg”, “nice”, “decor”, “hotel”, “close”, “hop”, “bu”, “line”, “well”, “lo”, “rambl”, “?”, “complaint”, “light”, “bathroom”, “poor”, “put”, “make”
Traveled on business	Beuatiful friendliest staff. great location just off the Ramblas, so less noise, great rooms, - all superiors & up have either a terrace or balcony, and its right across the street from Gaudis family estate home which is in the process of being transformed into a museum of sorts for Gaudi.	“beuati”, “friendliest”, “staff”, “great”, “locat”, “rambla”, “le”, “nois”, “great”, “room”, “superior”, “either”, “terrac”, “balconi”, “right”, “across”, “street”, “gaudi”, “famili”, “estat”, “home”, “process”, “transform”, “museum”, “sort”, “gaudi”

Table 1. Results of text pre-processing including stop words and punctuation removal, lower-case conversion and stemming.

Figure 2 details how the top 50 most-used words are distributed across the four considered trip types. This figure reveals that, in general, all the words appear in the four classes of documents, which renders classification difficult as there are no clear terms specifically associated with any particular type of document. Moreover, this figure reinforces the idea of searching for specific topics rather than specific words.

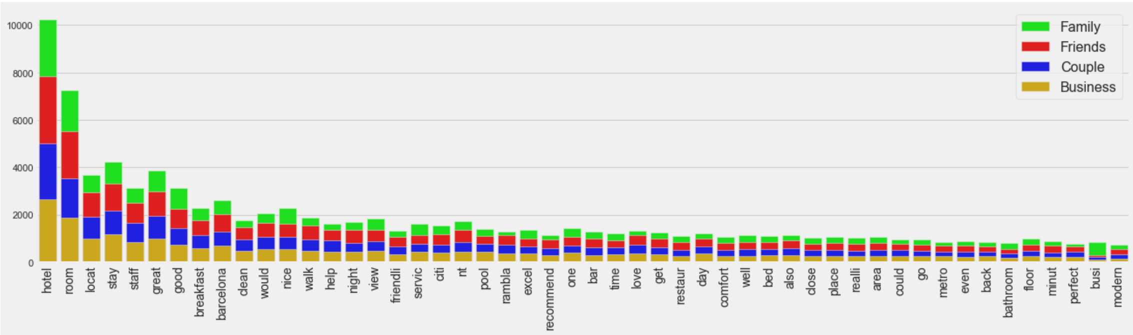


Figure 2. Word frequencies of the top 50 words per Trip type.

Topics analysis requires the codification of documents to enable the evaluation of any similarities. To answer RQ1, this research aims to determine the optimum type of document encoding capable of distinguishing between the four predefined classes in order to extract the specific topics associated with each. In this paper, we will use and compare three different methods of document encoding construction: LDA features, word2vec and Convolutional Encoding. In all three cases, the encoding will be defined as vectors of dimension 120.

LDA features are obtained by applying the LDA topic approach. The *Gensim* library (Řehuřek and Sojka, 2010) in Python provides some tools to predict topic distribution for each document. More specifically, *ldamulticore* is a parallelized Latent Dirichlet Allocation able that harnesses the power of multicore CPUs. As the number of topics was predefined as 120, the resulting vector for each document is probability distribution over this number of topics.

The second encoding method is based on word2vec representation. Again, the library provides an implementation of word2vec using skip-gram architecture. Vector size was set to 120 and negative sampling was selected with a value equal to 5. Document encoding is obtained by averaging the word2vec vectors of words.

The last encoding method is based on the concatenation of three 1D-CNNs built as a layer of a neural classifier able to predict the class of each document. The first 1D-CNNs consists of 40 filters of size 1, which means that documents are analyzed on the basis of unigrams; the second 1D-CNN consists of 40 filters of size 2, which refers to bigrams; and finally, the last 1D-CNNs are 40 filters of size 3, that is, trigrams. The concatenation of these three layers leads to the final encoding of size 120.

The three previous encodings were first compared in terms of accuracy when classifying the documents. In the case of LDA features and word2vec, the encoding is first calculated and then the classifier is trained and tested. The case of the neural encoding is different because the encoding is calculated once the classifier has been trained. This is an important difference, as the encoding is calculated with the general aim of maximizing the accuracy of the classifier. In other words, the convolutional neural encoder is not only a dense representation that captures semantic similarity, but it also reflects the differences between the considered classes.

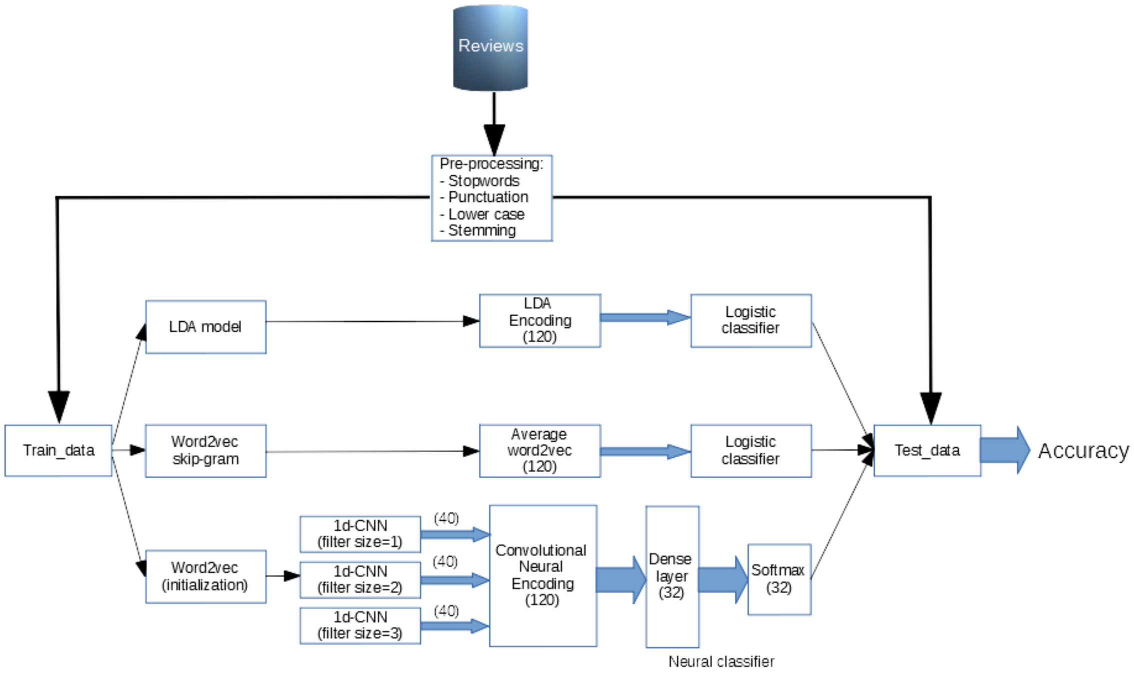


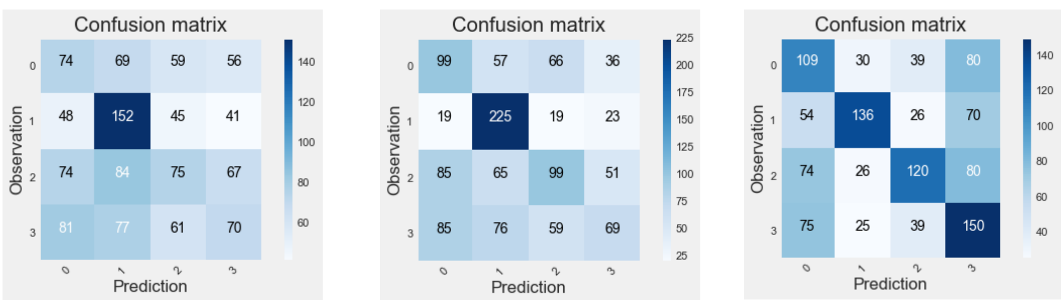
Figure 3. Block diagram of the proposed methodology.

Figure 3 shows the block diagram of the three document encodings and their classification performance. First, the four pre-processing stages, stopwords removal, punctuation removal, lower-case conversion, and stemming are applied to the original dataset. Next, the original dataset is split into two subsets, the training dataset (80%) and the test dataset (20%). The training data are used to train the classifiers for the three different document encodings. Hence, the training step defines the weights of the three classifiers. In the case of the convolutional neural encoder, Word2Vec provides only the initial values of the word embedding, as the embedding matrix and the convolutional neural encoding are also weights that will be fitted during the training time. The performance of the three classifiers is then calculated using the test dataset. Documents belonging to the test dataset are also encoded using the three encoding options and applied as inputs to the fitted classifiers.

5. Results

The accuracy of the three classifiers along with the confusion matrices are reported in Figure 4. The confusion matrix describes the performance of the classification model. Correct predictions are situated in the diagonal of the matrix while off-diagonal elements indicate the cases where the prediction differs from the ground truth (incorrect predictions). Classification accuracy is the ratio of correct predictions to total predictions. As shown in Figure 4, average word2vec and convolutional neural encoder clearly outperform LDA features classification performance. Although the classifications of the first two methods are similar, it is worth noting that in the case of the convolutional neural encoder, correct classifications are balanced, while in the case of average word2vec they are clearly unbalanced toward class 1, which represents travel for business.

LDA features | Average word2vec | Convolutional neural encoder

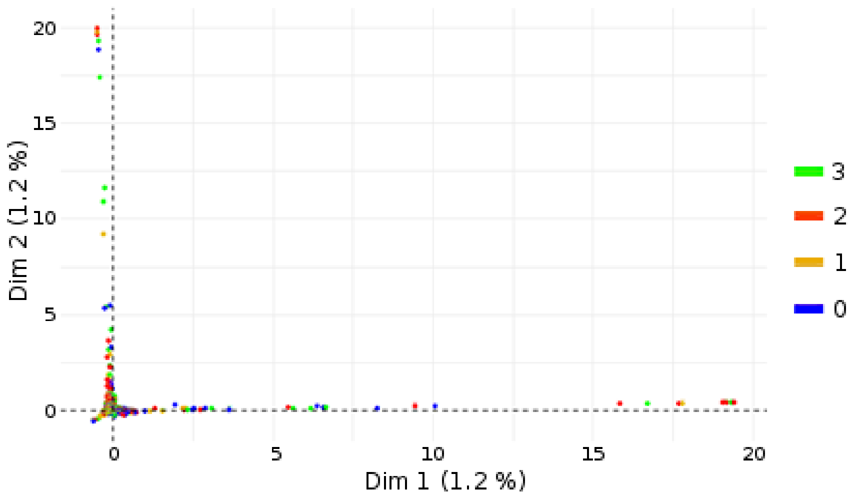


Accuracy: 0.3274 | Accuracy: 0.4288 | Accuracy: 0.4545

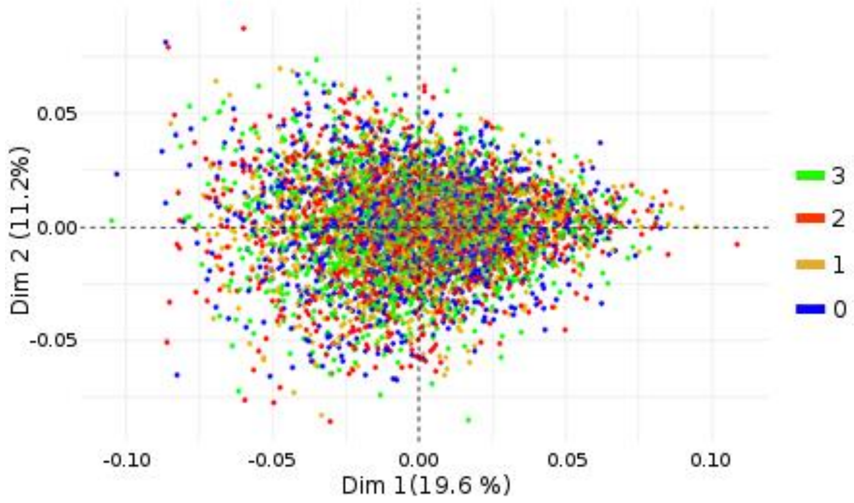
Class definition: 0-Couples; 1-Business; 2-Friends; 3-Family

Figure 4. Confusion matrices for the three encoding methods and classifiers.

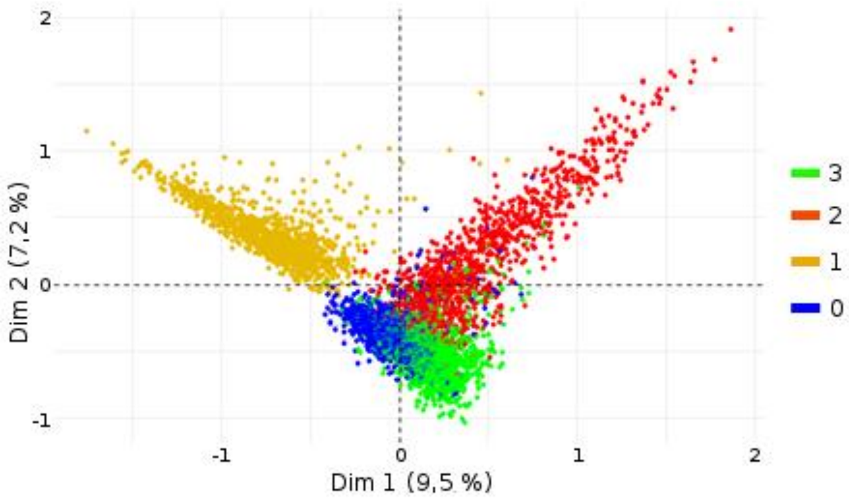
However, the main advantage of the convolutional neural encoder is that the vector values are also fitted as part of the classifier training. Hence, the obtained document encoding does not only show the semantic similarities between groups of documents but also emphasizes the dissimilarity to documents belonging to different classes. This property can be visualized using a correspondence analysis, which shows the correspondence between the items of the two basic categories, classes, and attributes according to their distance from each other (Greenacre & Blasius, 2006). Figure 5 depicts the bi-plot for the three proposed encoding methods. Each point represents a document and the color indicates the class that the document belongs to, with 0-blue Couples; 1-yellow Business; 2-red Friends; and 3-green Family.



(a) LDA features



(b) Average word2vec encoding



(c) Convolutional neural encoding

Figure 5. Correspondence analysis for the three proposed encodings.

The bi-plot for LDA features (Figure 5 (a)) distributes the documents along the orthogonal axes. Most are located next to the origin and there is a high dispersion of documents along the axes. Figure 5 (b) depicts the bi-plot for average word2vec encoding. It can be seen that there is no clear separation between classes. Finally, Figure 5 (c) shows the result for the convolutional neural encoding, where classes are clearly separated, as stated in H1. This is due to the encoding having been fitted as part of the classifier training, so the encoding not only collects the semantic similarities of documents but also emphasizes their semantic differences depending on the class they belong to. Hence, convolutional neural encoding is the best option for obtaining the topics that are specifically related to one travel type and no other.

For this purpose, when using convolutional neural encoding, the clustering k-means algorithm was separately applied to the set of reviews belonging to each class. The value of k was selected using the elbow criterion. Following this criterion, the heterogenicity is calculated from the sum of squares of the distances to centroids, and the elbow of the curve is considered to be the optimum value of k. Hence, 2, 3, 4, and 2 clusters were suggested for business, couples, family, and friends, respectively.

The most representative terms for each cluster were selected by using the so-called unique attributes (Toral et al., 2018), which in this case are the terms uniquely associated with one particular topic out of all the possible topics. The main advantage of using this selection is that unique attributes are the words with the best discriminative properties.

The unique attributes were identified by using an ANOVA test to compare the clusters term frequencies–inverse document frequencies (TF-IDF), which are the normalized values of term frequencies. TF-IDF balances the frequency of words with their infrequency in the document set. A posthoc Tukey's Honestly Significant Difference (HSD) test was then applied to associate an attribute with a specific cluster (topic) in the remaining topics related to each class.

Table 1, 2, 3 and 4 show the homogeneous topics found for couples, business, friends and family, trips respectively, along with their unique attributes, and a few selected examples per topic taken from the dataset. Some issues exist that are usually important to any type of hotel guest (bed comfort, breakfast quality, etc.). However, in this paper, convolutional neural encoding was fitted as a part of classifier training, so it emphasizes the semantic differences between the different trip types, that is, the topics that are more related to one travel type and no other. Thus, although words like "breakfast" or "bed" appear in many reviews, they are not the unique attributes of any trip type.

Regarding the **business trip reviews** (Table 2), the clustering analysis and the selection of unique attributes suggest two main topics. On the one hand, these reviews emphasize the experience of attending professional conferences, exhibitions, and meetings, including the business facilities provided by the hotel and its location with respect to the downtown area or another meeting venue (“attend”, “company”, “confer”, “congress”, “event”, “meet”, “venu”, “ride”, “downtown”). As can be observed, there are also attributes in this topic associated with the “Mobile World Congress” (“mobil”), the main world trade show devoted to the mobile communications industry. The venue where this event is held annually is the “Fira de Barcelona” (“fira”), one of the most important trade fair institutions in Europe, which hosts major events and receives thousands of professional visitors.

On the other hand, business trip reviews are specially focused on aspects related to the hotel environment and the room facilities and appliances. The time that these travelers stay in their rooms is probably longer than in the case of other more leisure-oriented travel types. Business guests often need to use their own rooms as a workplace and, in general, their need for rest and relaxation is greater. In addition, they usually travel alone. For this reason, business trip reviews highlight environmental factors such as illumination, noise, furniture, and general room comfort (“light”, “loud”, “corridor”, “door”, “glass”, “towel”, “mirror”, “rain”, “shower”, “bathroom”, “menu”). Furthermore, compared to other trip types, business travelers emphasize baggage handling and storage (“bag”) and sufficient space for clothes (“cloth”), as well as the presence of a TV in the bedroom and the TV channel offer (“TV”, “channel”).

Topics/Attributes	**Examples**
Experience of attending professional meetings and conferences:	“We were 200 people from the same company attending a conference, in the morning it was awful to use the lift as everybody had to use it!” “I was there for business. My company had an event not far from this place (Fira de Barcelona)”

Topics/Attributes	Examples
“attend”, “company”, “confer”, “congress”, “event”, “meet”, “venu”, “ride”, “downtown”, “Mobil”, “Fira”.	“The business facilities are excellent with fully functional meeting rooms and a plenary room with integrated systems and capacity for larger numbers of delegates” “The venue location is excellent as it was only 5 mins to the CCIB and walk-able into the old town” “Property is lovely and luxurious. However, it is a long train ride to downtown” “I stayed here with customers during Mobile World Congress. OK to stay for business purposes during the conference due to its proximity to the Fira”
Hotel environment and room comfort, facilities and appliances: “light”, “loud”, “corridor”, “door”, “glass”, “towel”, “mirror”, “rain”, “shower”, “bathroom”, “menu”, “TV”, “channel”, “bag”, “cloth”	“Room nicely decorated and equipped (b&o sound system, flat screen tv). Bedding is comfortable. Bathroom is also comfortable (bath+ shower) with good appliances. Good lightening and controls throughout the room” “For a 5-star hotel I wouldn’t expect damp smelling corridors, but on our floor there were maintenance men” “Thankfully the pool isn’t open on the ground level. I can’t imagine how much louder it would have been. Have sliding door closed and air on and can still hear everything outside the room”. “Air conditioning - in the morning, the sun was straight and in front of my windows. The glass of the window was really hot, and the room too. It was really unpleasant” “The towels and bed linens were luxurious! Towels thick and soft. A frosted door separates the bedroom from the bathroom suite and another frosted door separates the toilet from the rest of the bathroom”

Topics/Attributes	Examples
	"No mirror in the main room area to dry your hair and do make-up at natural light" "The room has a modern feel with dark furniture and lots of glasses, flat screen TV, stocked fridge, remote controlled curtains, rain shower…" "Bed extremely comfortable and a pillow menu is available if needed" "Good sized wardrobe to hang clothes and place to put suitcase. "The staff there was fantastic! Helping me to carry my bags" "TV selection is very poor, about 70 channels mostly local or international news channels, with no movie or sports channels"

Table 2. Topics, unique attributes, and examples of business trip reviews

Table 3 shows three main topics for the **couple trip reviews**. First, it is possible to distinguish some unique attributes related to couple celebrations in the reviews associated with this trip type ("anniversary", "birthday", "honeymoon", "partner"). Sometimes celebrating couples receive gifts from the hotel which are highly appreciated in their reviews (such as room upgrades, bottles of wine, cakes, etc.).

Second, Table 3 shows many unique attributes other than accommodation associated with the use of services and experiences provided or advised by the hotel and their prices. Compared to business trip reviews, couple trip reviews emphasize more aspects related to the global experience of visiting Barcelona and numerous leisure activities, both inside and outside the hotel. These activities and services generate many contacts and interactions with the members of staff ("member", "call", "phone", "tip", "order", "sent"). Sometimes these interactions are related to activities outside the hotel organized with the help of staff members or recommended by them, which is highlighted in the reviews: buying tickets for events, tours and cultural attractions, and public transport (“ticket”), calling for a taxi or booking local restaurants (“call”, “phone”), etc. On other occasions they are related to the services provided by the hotel itself, such as calls to room service or dinners, shows, and drinks in the hotel restaurants and lounges, often on Saturday evening (“even”, “Saturday”). Other highlighted services apart from accommodation per se are turn-down, bottled water, mini bar, and iron and ironing board (“turn”, “water”, “mini”, “iron”, “board”). Furthermore, unlike business travelers, couple reviews frequently refer to service prices and some unique attributes are related to their cost and payment (“pay”, “card”, “euro”). Finally, compared to business reviews, couple trip reviews try to connect more with other guests and refer to previous reviews related to these services posted on TripAdvisor (“state”), giving advice or a warning about some possible problems (“tip”, “opinion”, “avoid”, “wait”) and strongly approving or vehemently criticizing how the services were provided (“world”, “worst”).

Thirdly, couple trip reviews especially emphasize praise for the hotel staff’s work and helpfulness. As mentioned, couple reviews frequently refer to services and leisure experiences apart from accommodation that are provided or recommended by the hotel and generate many interactions with staff members. Some unique attributes are focused on highlighting the friendliness and help provided by staff members (“friendli”, “great”, “help”, “clean”).

Topics/Attributes	**Examples**
Couple celebrations: “anniversary”, “birthday”, “honeymoon”, “partner”	“For my partner’s birthday we were upgraded to a junior suit and had a birthday cake brought to the room which was such a lovely touch” “We were celebrating my husband´s birthday and our 15th anniversary during our stay. They gave us an upgraded room with lots of natural light and with great views of the city and partial view of port. On my husband´s birthday they delivered a Cava bottle and chocolate to our room, which I thought was very thoughtful”

Topics/Attributes	Examples
	"We checked in to this amazing hotel for 4 days during our honeymoon in Barcelona.it was an amazing experience. The hotel staff upgraded us to a sea view room which was amazing"
Services and experiences provided or advised by the hotel, beyond accommodation: "member", "call", "phone", "tip", "order", "sent", "ticket", "Saturday", "even", "turn", "water", "mini", "iron", "board", "pay", "card", "euro", "state", "opinion", "avoid", "wait", "world", "worst"	"From the front desk to room service and the breakfast buffet, staff members were very attentive, polite and accommodating" "Staff spoke good English and were extremely helpful; especially the concierge who called round local restaurants to find a local seasonal delicacy we wanted to try!" "We thought, we can phone room service to book some early morning tea" "Wonderful hotel - lovely staff. They smiled! Nothing was too much trouble and they gave us some wonderful tips and booked local restaurants" "(Restaurant) staff with very poor English, don't understand what you want, bringing things that you didn't order. " "The consommé and fresh lemon in hot water we ordered from room service was a tremendous help in our recovery" "She (staff member) was so kind - provided us with a doctor from their 24 hours on call service…In addition, she sent us water and chocolates to our room". "The staff was helpful and friendly and even assisted us in printing tickets for the Sagrada Familia at the ATM across the street" "The front desk was always helpful, from the offer at check-in to help us search for "El Clásico" tickets and the recommendation for a bar to watch it in otherwise, to calling our cab at 4am for our departure"

Topics/Attributes	Examples
	"The restaurant at the Hotel was phenomenal… impressive dinner with friends on a Saturday night when we struggled to get reservations elsewhere" "It was a wonderful experience, and we deliberately stayed over a Saturday evening to experience the opera singers' performance in the lounge" "The next night, no treats, no complimentary water, not even with turn-down service" "The mini bar is well stocked, but a little pricey, but I guess it is excepted from a 5 Star Hotel" "You should request a room view a view over the front. You will also need to request an iron and ironing board" "We also felt that having to pay 14 euros to use the gym (per day) was unnecessary and should have been included" "Breakfast at €25 was a tepid affaire. They charged my credit card as soon as we arrived" "Like one other guest has stated, it was hard to believe they charged for luggage storage" "Some reviewers went over-the-top, in my opinion, on their review of this buffet. It was nice, not amazing." "Just a tip: try to avoid the breakfast. It is 12€ per person and is just fee quiches. Around the hotel you can pay 12 for 2 persons a decent breakfast!!" "If you really have to take a taxi (highly not recommended) do not ask the doorman to call one for you as they'll add 3 Euros just for that, just wait 2 minutes and they'll come."

Topics/Attributes	Examples
	"The breakfast buffet, while pricey, was the best we had experienced anywhere in the world." "Our only recommendation is to improve the coffee and cappuccino (simply the worst we have ever had in a hotel)."
Praise for staff's work and helpfulness: "friendli", "great", "help", "clean"	"We were blown away by the staff's friendliness, professionalism, willingness to go out of their way to help in every way possible." "The hotel staff were great, very helpful and friendly." "The room was very comfortable and the cleaning service excellent, both daytime and turndown service."

Table 3. Topics, unique attributes, and examples of couple trip reviews

Table 4 shows the four main topics found in the **family trip reviews**. The first is related to this trip type's group nature. These reviews usually include the size of the family group and opinions about the suitability of the services provided for this size ("group"). On the other hand, getting around is more difficult for a group than it is for couples or individuals. Therefore, the proximity of the hotel to metro and bus routes ("rout") is significantly highlighted in family trip reviews. Furthermore, hotel cocktail bars are especially important because they are often used as a place to meet other members of the family ("cocktail"). These group activities also take place mainly at the weekend ("weekend").

A further topic is related to the attitude and manners shown by the hotel staff ("friend", "rude"). The frequency with which family trip reviews highlight this aspect is higher than it is in other trip types to the point of being one of their distinctive features. Unlike the couple trip reviews, which show unique attributes praising staff's work and helpfulness, family-trip reviews stand out for including many comments, both positive and negative, focused on the friendliness or rudeness of the staff, both in general and in the provision of specific services.

Table 4 also shows a number of unique attributes associated with hotel management and problem-solving. In these reviews, company policies ("polic") and the role played by managers ("manag") in solving different issues raised by guests ("explain", "happen", "told", "email") are highlighted. When the problems are solved, special mention is often made of the staff in charge ("mention"). Moreover, when the problems are solved with immediacy, this is often reflected in the reviews ("immedi"). Likewise, many reviewers focus on the mistakes made by the hotel ("mistak", "wrong"), such as incorrect charges and billing errors, booking mistakes, or wrong information given by the staff.

Finally, Table 4 shows some unique attributes associated with complimentary services, which proves their importance to these travelers. Family trip reviews highlight complimentary services provided by hotels ("provid") that are free of charge ("complimentar", "extra") such as toiletries, bowls of fruit, maps, baby amenities, phone chargers and adapters, and all kinds of gifts. Some of these services and possibly the most common in hotels have specific unique attributes associated with them such as bottles of wine or water, tea and coffee machines ("bottl", "tea", "machin"). On the other hand, family trip reviewers also highlight these kinds of services when they are not provided or involve an additional cost.

Topics/Attributes	**Examples**
Group experience: "group", "cocktail", "rout", "weekend"	"We were a <u>group</u> of 12 family members staying in Barcelona 2 nights." "There are 2 large patios on the roof…and our <u>group</u> used it several times to enjoy our wine and food." "We used the bar as a get together. Bar staff were fantastic, in particular David who was making <u>cocktails.</u>" "The roof restaurant/bar is great for a <u>cocktail</u> after a day of touring and offers snacks and fantastic views."

Topics/Attributes	Examples
	"The big plus is that it is very close to the Clot Metro stop on the Red Line 1 Metro route." "As a family with three kids this was ideal. A modern apartment…ideally located on bus route." "Booked this for a weekend away with family. Out of town but very close to the Metro so it was easy and cheap to get around."
Staff´s attitude and manners: "friend", "rude"	"The staff was kind, friendly and attentive." "Staff at front desk was not friendly which really ruins your happy vacation mood. One of the staff…was exceptionally rude." "When asked by late check out, the clerk simply read what was written…, ignoring our question and replying in very rude manners." "I did not encounter any rudeness nor arrogance with any of the hotel staff. They were courteous and helpful."
Hotel management and problem-solving: "polic", "manag", "explain", "happen", "told", "email", "mention", "immedi", "mistak", "wrong"	"I am shocked to hear from the management that they cannot do it...They did not change the bookings and forfeited all our money. When I requested them, they told me its company policy." "We arranged to meet the manager of the hotel the next day to discuss what had happened and to be compensated." "I called down to reception to see if we could move…we were told it was possible, but it would cost 100 euros/night + tax, which I explained I did not understand if it was the same category." "Prior to traveling anywhere, I always send an email ahead to the general manager, explaining our situation…I received an email stating that every one of our dietary needs would be met."

Topics/Attributes	Examples
	"The fellow at the desk was not friendly about their mistake and he acted as though he could not understand how this could have happened." "Despite complaints to the manager on duty, no response came to fore. They refused to share the name of employee who had cleaned the room for us to give police complaint." "Other members of staff I must mention who went above and beyond to be on hand during a truly difficult time were Javier in the breakfast restaurant and Diego on reception" "We had a minor issue with our bathroom which was immediately attended." "I explained that we did not ate any breakfast at the hotel, then the concierge confirmed that it was hotel's mistake." "The experience throughout went from bad to worse: wrong information given at reception regarding restaurants available."
Complimentary services: "provid", "complimentar", "extra", "bottl", "tea", "machin"	"They provide a good kettle with coffee, tea and sugar." "Staff are extremely helpful and provide you with maps to get around." "Along with the baby cot they also provided us with a baby tub and baby organic toiletries." "Numerous bathroom amenities were also provided which was a lovely surprise." "Everyday we had complimentary water, beautiful turndown…, we received a bottle of wine, cupcakes, the extras never appeared to stop."

Topics/Attributes	Examples
	"There was complimentary champagne on ice and chocolates. Bottled water was complimentary and replaced on a daily basis." "As most reviews have shown the spa/gym is a separate building and incurs extra charges, which is not worth it." "The hotel does not provide tea/coffee making amenities and does not supply bottled water." "There are fewer complimentary amenities given to guests than I've experienced in other five star hotels…A Nespresso machine is available in the room but the pods are charged on your bill"

Table 4. Topics, unique attributes, and examples of family trip reviews

Regarding the **friends trip reviews** (Table 5), the clustering analysis suggests two topics. The bi-plot for convolutional neural encoding (Figure 5 (c)) shows that friends trip is the class with a higher dispersion of documents along the axes. The main reason for this is that this TripAdvisor category includes reviews written by guests in groups made up of friends who are traveling without any children or partners, couples who travel with other friends, and some families who travel with other families in a larger group. Due to this heterogeneity, there are fewer unique attributes in friends trip reviews.

First, Table 5 shows one unique attribute associated with the nightlife in the hotel and in the surrounding area (night), which in general terms is more important to these travelers. Second, it is possible to distinguish some unique attributes associated with stays in suites instead of regular hotel rooms. This kind of accommodation (suite, junior suite) usually has more space than a typical hotel room, connects two or more rooms, and facilitates interaction within the group of friends, which is appreciated by this type of guests. The use of the sofa is highlighted in this type of accommodation, generally as a bed (sofa), which helps reduce costs. Also, the reviews reveal a greater use of room service in these suites (deliv).

Topics/Attributes	Examples
Nightlife: "night"	"The rooftop bar was the perfect place to end the night." "The restaurant and bar are great, as is the nightclub at the top." "Royalty treatment and lots of nightlife fun!"
Stay in suites/shared rooms: "suit", "junior", "sofa", "deliv"	"The suite was very spacious and luxurious" "We were six people and we had two corner suites for 4 nights. For us this was a perfect place to stay in Barcelona." "We had a junior suite which was roomy enough for us." "The bed was very comfortable…even the pull out sofa (bed 2) which was separated by a partition that made it very private, was comfortable (for a sofa bed)." "beds were amazingly comfortable, even the sofa bed for the 3rd person!" "We ran up to reception and explained the situation, they said you can order room service and have it delivered anywhere in the hotel, great!" "Room service meal was tasty, delivered promptly and reasonably priced."

Table 5. Topics, unique attributes, and examples of friends trip reviews

6. Discussion and implications

6.1 Comparison with state-of-the-art techniques and theoretical contributions

The identification of consumer preferences using natural language processing is an emergent topic due to the high volume of information available on the web. Managers and practitioners can take advantage of the open-ended opinions shared on websites to determine the main topics of interest as well as the strengths and weaknesses of their products and services. This is of particular interest for the tourism sector, where websites such as TripAdvisor have become a standard source of information for travelers. Despite the heterogeneity of potential users, they can be categorized a priori by some criteria, with the trip type being one of the most used criteria in the previous literature. However, and although reviews can be easily separated according to the trip type, it is challenging to determine the customer preferences associated with each trip type. This is mainly because many reviews address similar issues that apply across all types of hotel guests. The focus of this paper is, precisely, on determining the unique topics of interest given a set of predefined classes, such as those represented by the trip types. Therefore, this work's main contributions to the existing literature are:

First, our study identifies unique topics associated with a set of predefined classes and improves on the results provided by traditional clustering methods that follow an unsupervised learning scheme. LDA is still the most frequently used method for topic modeling. However, in its basic formulation, it is not capable of modeling complex data relationships such as correlations between topics or their temporal patterns (Vayansky, & Kumar, 2020; Curiskis et al., 2020). Some topic models dealing with advanced topic relationships include the correlated topic model (Blei & Lafferty, 2006), the Pachinko allocation model (Li & McCallum, 2006), and spatio-temporal information management (Asghari et al., 2020). More recent works have modeled more complex data correlation relationships by making use of neural networks, for example, variational autoencoders (Zhao et al., 2021) and generative adversarial networks (Wang et al., 2019) using an LDA encoding of documents. In contrast with previous approaches, this paper searches for non-correlated topics, given that the documents belong to a predefined number of classes. The identification of the topics of interest when consumers are pre-classified as belonging to a specific class results in some redundant topics between classes. Hence, the clustering of topics does not provide evidence about the specific preferences of each specific group. To overcome this issue, this work provides a methodology for collecting the distinctive topics of interest for each class that enables the specific preferences per class to be determined.

Second, this paper utilizes convolutional neural encoding as the mathematical representation of documents to be used with the k-means algorithm. Unlike other encoding techniques, we have proven that convolutional neural encoding fitted in trained classifier groups separates documents belonging to the same class, facilitating the identification of the distinctive unique topics. From a methodological viewpoint, our main contribution is the novel encoding of documents using convolutional neural encoding fitted as part of a classifier. This encoding includes the distinctive feature of each class, so the resulting clustering algorithm can provide not only the main topics of interest for that class but also the distinctive topics in relation to other classes included in the classification task. As posited by H1, convolutional neural encoding was demonstrated to provide the best discrimination between classes compared to other encoding techniques. The main novelty here is the idea of using the neural network encoding fitted during the classifier training time. Some other neural network architectures that exploit the discovery of local features invariant to translation and the sequential order of text could also be used, including LSTM, the CNN-LSTM hybrid schemes (Behera et al., 2021), or even including some other advanced features such as attention mechanism or ensemble learning (Liang & Yi, 2021).

Third, this study identifies unique topics for different traveler profiles, allowing hotel managers to focus on a specific customer segment. As posited by H2, it is possible to find the specific topics that are related to each particular travel type and no other. Previous studies have only used the variable trip type as a control variable, so only some specific issues have been studied as being impacted by trip type selection, i.e., travel expenditure (Sung et al., 2001), agent destination recommendations (Klenosky & Gitelson, 1998), specific hotel features (Chu & Choi, 2000), revisiting (Hu et al., 2019b), and expectations (Liu et al., 2013). Our study shows the homogeneous topics found for business, couples, friends, and family trips and also their unique attributes.

6.2 Practical implications

The findings of this study also offer several practical implications. First, as user-generated content and eWOM play important roles in influencing travelers’ decisions, they must be strictly monitored by hotel managers and tourism practitioners. Although there are some issues that are usually important to every traveler, the convolutional neural encoding of reviews shows that there are topics and unique attributes that are more related to one of the travel types and not to the others. Therefore, hotels must reformulate their offer and marketing strategies based on the type of travelers they cater to. Better specialized knowledge of guests' expectations and requirements allows managers to focus on the important issues to provide a better service and optimize resources and traveler satisfaction. For example, for business travelers, hotel environment and general room comfort should be specially analyzed and designed as they usually have a greater need for rest and they often use the room as a workspace. Couples highlight their special celebrations (anniversary, honeymoon, etc.), so they greatly appreciate any complimentary upgrades and presents related to these events. They also highlight many services and experiences apart from accommodation that generate a large number of contacts with staff members. As this type of guest very much appreciates any help or tips for leisure activities both inside and outside the hotel, staff members should be trained and have the necessary resources to provide these specific services. In the case of family travelers, hotels should carefully plan the suitability of the services and meeting places provided for groups, which can be large and include children, and the proximity to public transport routes. The comments about situations and problems to be solved by hotel managers and staff are usually longer for family trips than they are for the other trip types, so managers need to be prepared for them. In this context, family travelers especially emphasize the role played by managers in solving these issues and the attitude and manners shown by the hotel staff. Besides, complimentary services provided by hotels are truly important for these guests and are often highlighted in their reviews. Finally, for friend travelers, aspects such as nightlife in the hotel and the surrounding area are usually more important and should be managed by tourism practitioners.

Second, as traveler expectations and demands change depending on the trip type, hotels have to make a serious effort not only to adapt their services to suit these differences but also to work with proper tools to analyze them and detect any changes over time. To fulfill this objective, this paper offers a methodology that identifies the unique topics and attributes that belong to each of the considered trip types.

Third, this research can also have implications for website management and recommendation systems. For example, depending on the profile of the user searching for information, websites such as TripAdvisor can display the reviews associated with the travel type's specific topic of interest first and even recommend options with high ratings for these topics.

6.3 Limitations and further research

This paper relies on neural encoding provided by a convolutional neural network. However, neural networks have many hyperparameters that must be chosen prior to the training step. In this case, we chose a neural encoding dimension of 120. Although this value is high enough to capture the features related to the semantic meaning of documents, different values could have been tested. Additionally, convolutional neural networks are not the only encoding technique available in deep learning. LSTM is a recurrent neural network that can also be used for encoding documents. The main difference with respect to convolutional neural networks is that they can identify long-term dependencies within text. Finally, autoencoders, which can be built based on convolutional or LSTM networks, are another option.

Regarding the dataset, four trip types categories were selected. There are some others but they exhibit a certain degree of overlapping with the chosen types. For example, the solo trip type is quite similar to the business trip type, as many business travelers usually travel alone. Therefore, and for the sake of clearly differentiating between classes, the business trip type was chosen and the solo trip type discarded.

This work could be extended by comparing different encoding schemes based on convolutional and LSTM networks. Although the Skip-gram model was used for word representation, there are some other word embedding models that could also have been tested such as GloVe (Global Vectors) and ELMo (Embeddings from Language Models). Finally, findings can be compared to those obtained for a different dataset.

7. Conclusions

This paper addresses the identification of the unique topics of interest associated with four different traveler profiles according to their trip type. The proposed methodology relies on the convolutional neural encoding of documents, which is able to capture the features that semantically distinguish documents belonging to different classes. From a methodological point of view, the paper demonstrates that the neural encoding fitted as part of a classifier maximizes the discrimination of documents when compared to other encoding schemes. Findings reveal travelers' preferences depending on their trip type, enabling hotel managers to focus on the topics raised by customers to decide the best strategy to approach their target clientele.

References

Agarwal, B., Ramampiaro, H., Langseth, H., & Ruocco, M. (2018). A deep network model for paraphrase detection in short text messages. Information Processing & Management, 54(6), 922-937.

Ahmed, A., Toral, S., Shaalan, K., & Hifny, Y. (2020). Agent Productivity Modeling in a Call Center Domain Using Attentive Convolutional Neural Networks. Sensors, 20(19), 5489.

Alami, N., Meknassi, M., & En-nahnahi, N. (2019). Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. Expert systems with applications, 123, 195-211.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. S. (2020). A topic modeling framework for spatio-temporal information management. Information Processing & Management, 57(6), 102340.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems, 115, 279-294.

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Information Processing & Management, 58(1), 102435.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.Chu, R.K., Choi, T., 2000. An importance-performance analysis of hotel selection factors in the Hong Kong hotel industry: a comparison of business and leisure travelers. Tourism Management 21 (4), 363–377.

Blei, D., & Lafferty, J. (2006). Correlated topic models. Advances in neural information processing systems, 18, 147.

Cambria, E. (2016). Affective Computing and Sentiment Analysis. IEEE Intelligent Systems, 31 (2), 102-107.

Chu, R. K., & Choi, T. (2000). An importance-performance analysis of hotel selection factors in the Hong Kong hotel industry: a comparison of business and leisure travellers. Tourism management, 21(4), 363-377.

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management, 57(2), 102034.

de Lira, V. M., Macdonald, C., Ounis, I., Perego, R., Renso, C., & Times, V. C. (2019). Event attendance classification in social media. Information Processing & Management, 56(3), 687-703.

Dimanche, F., & Havitz, M. E. (1994). Consumer behaviour and tourism: Review and extension of four study areas. Journal of Travel & Tourism Marketing, 3(3), 37-57.

Ding, K., Choo, W. C., Ng, K. Y., & Ng, S. I. (2020). Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation. International Journal of Hospitality Management, 91, 102676.

Djaballah, K. A., Boukhalfa, K., & Boussaid, O. (2019). Sentiment Analysis of Twitter Messages using Word2vec by Weighted Average. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 223-228). IEEE.

Fathi, E., & Shoja, B. M. (2018). Deep neural networks for natural language processing. In Handbook of statistics (Vol. 38, pp. 229-316). Elsevier.

Fournier-Viger, P., Gomariz, A., Campos, M., Thomas, R., 2014. Fast vertical mining of sequential patterns using co-occurrence information. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 40-52.

Gao, B., Hu, N., & Bose, I. (2017). Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews. Decision Support Systems, 95, 1-11.

Gao, B., Li, X., Liu, S., & Fang, D. (2018). How power distance affects online hotel ratings: The positive moderating roles of hotel chain and reviewers' travel experience. Tourism Management, 65, 176-186.

Greenacre, M., & Blasius, J. (Eds.). (2006). Multiple correspondence analysis and related methods. CRC press.

He, J., Wang, X., Vandenbosch, M. B., & Nault, B. R. (2020). Revealed preference in online reviews: Purchase verification in the tablet market. Decision Support Systems, 132, 113281.

Hsieh, S., O'Leary, J. T., & Morrison, A. M. (1992). Segmenting the international travel market by activity. Tourism management, 13(2), 209-223.

Hu, F., Teichert, T., Liu, Y., Li, H., & Gundyreva, E. (2019a). Evolving customer expectations of hospitality services: Differences in attribute effects on satisfaction and Re-Patronage. Tourism Management, 74, 345-357.

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019b). What do hotel customers complain about? Text analysis using structural topic model. Tourism Management, 72, 417-426.

Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews–a text summarization approach. Information Processing & Management, 53(2), 436-449.

Iso-Ahola, S. E. (1982). Toward a social psychological theory of tourism motivation: A rejoinder. Annals of tourism research, 9(2), 256-262.

Jang, S., Bai, B., Hu, C., & Wu, C. M. E. (2009). Affect, travel motivation, and travel intention: A senior market. Journal of Hospitality & Tourism Research, 33(1), 51-73.

Ji, Y., Wu, W., Chen, S., Chen, Q., Hu, W., & He, L. (2020). Two-stage sentiment classification based on user-product interactive information. Knowledge-Based Systems, 106091.

Khorsand, R., Rafiee, M., & Kayvanfar, V. (2020). Insights into TripAdvisor's online reviews: The case of Tehran's hotels. Tourism Management Perspectives, 34, 100673.

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. Decision Support Systems, 134, 113302.

Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, 15-29.

Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021) Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply?. Tourism Management, 83, 104241.

Klenosky, D. B., & Gitelson, R. E. (1998). Travel agents' destination recommendations. Annals of Tourism Research, 25(3), 661-674.

Kotler, P., Bowen, J., & Makens, J. (1998). Marketing for hospitality and tourism (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kumar Behera, R., Kumar Rath, S., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2019). Distributed centrality analysis of social network data using MapReduce. Algorithms, 12(8), 161.

Kumari, A., Behera, R. K., Sahoo, K. S., Nayyar, A., Kumar Luhach, A., & Prakash Sahoo, S. (2020). Supervised link prediction using structured-based feature extraction in social network. Concurrency and Computation: Practice and Experience, e5839.

Lahlou, F. Z., Benbrahimand, H., Mountassir, A., & Kassou, I. (2013). Context extraction from reviews for Context Aware Recommendation using Text Classification techniques. In 2013 ACS International Conference on Computer Systems and Applications (AICCSA) (pp. 1-4). IEEE.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495-540.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).

Li, M., & Cai, L. A. (2012). The effects of personal values on travel motivation and behavioral intention. Journal of Travel Research, 51(4), 473-487.

Li, S., Chua, T. S., Zhu, J., & Miao, C. (2016). Generative topic embedding: a continuous representation of documents. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 666-675).

Liang, D., & Yi, B. (2021). Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification. Information Sciences, 547, 271-288.

Liu, S., Law, R., Rong, J., Li, G., & Hall, J. (2013). Analyzing changes in hotel customers' expectations by trip mode. International Journal of Hospitality Management, 34, 359-371.

Lockyer, T. (2005). Understanding the dynamics of the hotel accommodation purchase decision. International Journal of contemporary hospitality management.

Luo, J. M., Vu, H. Q., Li, G., & Law, R. (2020). Topic modelling for theme park online reviews: analysis of Disneyland. Journal of Travel & Tourism Marketing, 37(2), 272-285.

Ma, Y., Xiang, Z., Du, Q., & Fan, W. (2018). Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. International Journal of Hospitality Management, 71, 120-131.

Martínez-Torres, M. R., & Diaz-Fernandez, M. C. (2014). Current issues and research trends on open-source software communities. Technology Analysis & Strategic Management, 26(1), 55-68.

Martinez-Torres, M. R., & Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. Tourism Management, 75, 393-403.

Martínez-Torres, M. R., Arenas-Marquez, F. J., Olmedilla, M., & Toral, S. L. (2018). Identifying the features of reputable users in eWOM communities by using Particle Swarm Optimization. Technological Forecasting and Social Change, 133, 220-228.

Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. International journal of hospitality management, 34, 99-107.

Merlo, E. M., & de Souza Joatilde, I. (2011). Consumers attribute analysis of economic hotels: An exploratory study. African Journal of Business Management, 5(21), 8410-8416.

Meškelė, D., & Frasincar, F. (2020). ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. Information Processing & Management, 57(3), 102211.

Middleton,V.T. (1994). Marketing in travel and tourism (2nd ed.). Avon, UK: Bath Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv: 1301.3781.

Mitra, S., & Jenamani, M. (2021). Helpfulness of online consumer reviews: A multi-perspective approach. Information Processing & Management, 58(3), 102538.

Moro, S., Lopes, R. J., Esmerado, J., & Botelho, M. (2020). Service quality in airport hotel chains through the lens of online reviewers. Journal of Retailing and Consumer Services, 56, 102193.

Morrison, A. M. (1996). Hospitality and travel marketing (2nd ed.). Albany,NY: Delmar

Navío-Marco, J., Ruiz-Gómez, L. M., & Sevilla-Sevilla, C. (2018). Progress in information technology and tourism management: 30 years on and 20 years after the internet-Revisiting Buhalis & Law's landmark study about eTourism. Tourism management, 69, 460-470.

Nie, R. X., Tian, Z. P., Wang, J. Q., & Chin, K. S. (2020). Hotel selection driven by online textual reviews: Applying a semantic partitioned sentiment dictionary and evidence theory. International Journal of Hospitality Management, 88, 102495.

Nilashi, M., Ibrahim, O., Yadegaridehkordi, E., Samad, S., Akbari, E., & Alizadeh, A. (2018). Travelers decision making using online review in social network sites: A case on TripAdvisor. Journal of computational science, 28, 168-179.

Peng, H. G., Zhang, H. Y., & Wang, J. Q. (2018). Cloud decision support model for selecting hotels on TripAdvisor. com with probabilistic linguistic information. International Journal of Hospitality Management, 68, 124-138.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14 (3), 130-137

Radojevic, T., Stanisic, N., Stanic, N., & Davidson, R. (2018). The effects of traveling for business on customer satisfaction with hotel services. Tourism Management, 67, 326-341.

Rajaguru, R., & Hassanli, N. (2018). The role of trip purpose and hotel star rating on guests' satisfaction and WOM. International Journal of Contemporary Hospitality Management, 30, 5, 2268-2286.

Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. Retrieved from genism. org.

Ren, L., Qiu, H., Wang, P., & Lin, P. M. (2016). Exploring customer experience with budget hotels: Dimensionality and satisfaction. International Journal of Hospitality Management, 52, 13-23.

Ren, Z., Shen, Q., Diao, X., & Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. Information Processing & Management, 58(3), 102532.

Rhee, H. T., & Yang, S. B. (2015). How does hotel attribute importance vary among different travelers? An exploratory case study based on a conjoint analysis. Electronic markets, 25(3), 211-226.

Schuckert, M., Liang, S., Law, R., & Sun, W. (2019). How do domestic and international high-end hotel brands receive and manage customer feedback? International Journal of Hospitality Management, 77, 528-537.

Shen, Y. S., Choi, H. C., Joppe, M., & Yi, S. (2020). What motivates visitors to participate in a gamified trip? A player typology using Q methodology. Tourism Management, 78, 104074.

Sung, H. H., Morrison, A. M., Hong, G. S., & O'Leary, J. T. (2001). The effects of household and trip characteristics on trip types: a consumer behavioral approach for segmenting the US domestic leisure travel market. Journal of Hospitality & Tourism Research, 25(1), 46-68.

Teso, E., Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2018). Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. Technological Forecasting and Social Change, 129, 131-142.

Toral, S. L., Martínez-Torres, M. R., & Gonzalez-Rodriguez, M. R. (2018). Identification of the unique attributes of tourist destinations from online reviews. Journal of Travel Research, 57(7), 908-919.

Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. Tourism Management, 80, 104122.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

Vu, H. Q., Li, G., & Law, R. (2020). Cross-country analysis of tourist activities based on venue-referenced social media data. Journal of Travel Research, 59(1), 90-106.

Walls, A., Okumus, F., Wang, Y., & Kwun, D. J. W. (2011). Understanding the consumer experience: An exploratory study of luxury hotels. Journal of Hospitality Marketing & Management, 20(2), 166-197.

Wang, R., Zhou, D., & He, Y. (2019). Atm: Adversarial-neural topic model. Information Processing & Management, 56(6), 102098.

Wang, R., Ridley, R., Qu, W., & Dai, X. (2021). A novel reasoning mechanism for multi-label text classification. Information Processing & Management, 58(2), 102441.

Xu, X. (2020). How do consumers in the sharing economy value sharing? Evidence from online reviews. Decision Support Systems, 128, 113162.

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. International journal of hospitality management, 55, 57-69.

Yadav, M. L., & Roychoudhury, B. (2019). Effect of trip mode on opinion about hotel aspects: A social media analysis approach. International Journal of Hospitality Management, 80, 155-165.

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, 111-121.

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., & Zhuang, F. (2021) A neural topic model with word vectors and entity vectors for short texts. Information Processing & Management, 58(2), 102455.