Machine learning for assessing quality of service in the hospitality sector based on customer reviews

Vladimir Vargas-Calderón ·
Andreina Moros Ochoa ·
Gilmer Yovani Castro Nieto ·
Jorge E. Camargo

Abstract The increasing use of online hospitality platforms provides firsthand information about clients preferences, which are essential to improve hotel services and increase the quality of service perception. Customer reviews can be used to automatically extract the most relevant aspects of the quality of service for hospitality clientele. This paper proposes a framework for the assessment of the quality of service in the hospitality sector based on the exploitation of customer reviews through natural language processing and machine learning methods. The proposed framework automatically discovers the quality of service aspects relevant to hotel customers. Hotel reviews from Bogotá and Madrid are automatically scrapped from Booking.com. Semantic information is inferred through Latent Dirichlet Allocation and FastText, which allow representing text reviews as vectors. A dimensionality reduction technique is applied to visualise and interpret large amounts of customer reviews. Visualisations of the most important quality of service aspects are generated, allowing to qualitatively and quantitatively assess the quality of service. Results show that it is possible to automatically extract the main quality of service aspects perceived by customers from large customer review datasets. These findings could be used by hospitality managers to understand clients better and to improve the quality of service.

Vladimir Vargas-Calderón
Laboratorios de Investigación en Inteligencia Artificial y Computación de Alto Desempeño, Human Brain Technologies, Bogotá, Colombia.
Grupo de Superconductividad y Nanotecnología, Departamento de Física, Univesidad Nacional de Colombia, Bogotá, Colombia. E-mail: vvargasc@unal.edu.co

Andreina Moros Ochoa
Colegio de Estudios Superiores de Administración, Diagonal 35 #5a-57, Bogotá, Colombia.

Gilmer Yovani Castro Nieto
Departamento de Administración de Empresas, Pontificia Universidad Javeriana, Carrera 7 No. 40B-36, Bogotá, Colombia.

Jorge E. Camargo
System Engineering Department, Fundación Universitaria Konrad Lorenz, Bogotá, Colombia.

1 Introduction

Quality of service is a fundamental element in the hospitality industry, since delivering high quality of service is a differential feature that positively impacts hospitality businesses, particularly on their performance and competitive position (Wong Ooi Mei et al., 1999). The quality of service is a key ingredient that makes hospitality businesses thrive because there is a direct link between the quality of service and customer satisfaction, which translates into loyalty and repurchasing intention (Anderson and Sullivan, 1993; Ghotbabadi et al., 2012). Accordingly, the need for measurement of quality of service in the hospitality sector has been identified as critical in the last few decades because measuring it allows businesses to set goals related to the improvement of profitability, keeping track of their fulfilment.

In this order of ideas, frameworks to measure the quality of service have been built responding to different measuring needs. Researchers and managers usually disregard standardised frameworks since not every hospitality business thrives for the same population niche, meaning that failing to achieve a gold standard does not imply worse business performance. Instead, customer-oriented frameworks are preferred, as they directly correlate with customer satisfaction, which is related to the achievement of the customer's expectations (Anderson and Sullivan, 1993).

As might be expected, researchers and managers have produced a plethora of various methods to measure the quality of service, which have been applied to the hospitality sector (Ghotbabadi et al., 2012; Lai et al., 2018). The process of obtaining and gathering information from the clients has generally been done through surveys, which are time- and money-consuming. However, the growth of online hospitality platforms has generated firsthand data about customers that is not being used by these conventional methods. Understanding this data and finding ways to improve how customers perceive the quality of service is utterly essential, primarily because travellers use to write and read online reviews at any stage of their travel to make purchase decisions (Zhou et al., 2020; Pourfakhimi et al., 2020).

The development of these technological platforms enables the possibility to share opinions and assessments between users of services or products, generating a large volume of valuable information for the sector, which researchers can access to improve quality of service models (Xiang et al., 2017). Through the Internet, consumers actively use the dissemination of information by engaging in social dynamics where they receive and emit individual interactions, comments and decisions. This engagement is also known as the electronic word of mouth (eWOM) (Williams et al., 2019; Pourfakhimi et al., 2020). In fact, it has been shown that potential tourists are more likely to seek information about a destination through messages or comments from other consumers (Abubakar et al., 2017) in virtual platforms to make purchasing decisions (Vallejo et al., 2015; Septianto and Chiew, 2018; Kim et al., 2018). Therefore, eWOM is important because it improves tourists' experiences before, during, and after the trip (Rahmani et al., 2018). Those customer opinions and assessments about the quality of service are user gathered content that is valuable for managers to understand customers better and modify internal processes to improve user satisfaction.

The amount of information produced in these platforms sets a Big Data scenario, which –in recent years– has played a central role in tourism and hospitality research and delineates how it might evolve in the future (Mariani et al., 2018;

Mariani, 2019). The use of Big Data might allow making strategic decisions using current customer data (Lamest and Brady, 2019) and improve the competitiveness of tourism organisations and destinations (Buhalis, 2019). Most of this Big Data is expressed in text or images (customer reviews). It has been proposed that such Big Data can be exploited through artificial intelligence methods (Buhalis et al., 2019) able to process text (Xiang et al., 2019; Ma et al., 2018a) and images (Ma et al., 2018b) to assess several aspects of the hospitality industry. This is of particular usefulness, taking into account other sources different from online hospitality platforms such as social media (Moro et al., 2019; Lin et al., 2020).

The availability of this Big Data is crucial, especially to better assess the quality of service in the hospitality sector using customer reviews (Rahmani et al., 2019). Therefore, the ability to effectively analyse data, using, in occasions dedicated software, becomes a crucial aspect of hotel management. Consequently, it is natural to ask ourselves if we can measure and identify the main topics of the perceived quality of service in hospitality businesses from non-structured data contained in online customer reviews. In principle, the answer is yes: this is exactly what quality of service methods do. However, these methods would require investing human resources to extract meaningful and measurable statistics from the reviews so that the retrieved information can fit the dimensions and variables typically measured by the conventional quality of service measurement methods. In this work, we propose an alternative to assess the quality of service in the hospitality sector through the use of Natural Language Processing (NLP), a branch of artificial intelligence, to tackle the problem of inferring meaningful and measurable information from large customer opinions datasets that can be gathered from online hospitality platforms.

Specifically, we propose an NLP framework for rigorously studying customer opinions datasets, whose contribution is two-fold. First, our framework helps in qualitatively assessing the quality of service, focusing on and highlighting the main topics that concern the customers and the topics that comfort them. The fine-grained topic information can be obtained at different levels: city-, location- and business-level. Second, we provide an analysis of the identified topics and their relation to customers' final scores, which offers a quantitative measure of the quality of service for each topic.

We exemplify our NLP framework by obtaining city-level information about concern and comfort topics from hospitality customers in Bogotá, Colombia, and Madrid, Spain, through online reviews written in Spanish found at Booking.com for any hospitality business. We decided to study these two cities because they are capitals that generally have significant tourist traffic. Dhar (2015) highlights the importance of tourism development in emerging economies, as is the case of Colombia. Ministerio de Comercio, Industria y Turismo (2020) claims that in recent years, foreigner arrivals to the country have grown exponentially: in 2018, more than 4.2 million non-resident visitors, 8,504,778 travellers arrived in Bogotá on national flights and 4,465,741 on international flights; hotel occupancy in the country was 56.7%. For its part, Spain ranks third in the EU with the highest number of international tourists, 83.7 million non-resident travellers visited the country in 2019. 87.4% of these visits were for leisure, recreation and vacations and 6.4% for business and professional reasons (Instituto Nacional de Estadística, 2020). Specifically, Madrid recorded 10.4 million visitors in 2018 (Instituto Na-

cional de Estadística, 2020) and the country registered a hotel occupancy rate of more than 73%.

This paper is divided as follows. Section 2 shows an overview of machine learning applications in the hospitality sector. In section 3 we expose a deeper review of the methods used for text analysis in the hospitality sector. We also describe the data set used for this study and explain in detail the design of our research. Later, in section 4 we present the results of this work and finally conclude in section 5.

2 Literature review

Some of the main historical contributions to the quality of service assessment are the model proposed by Gronroos (1984), the widely used Servqual (Parasuraman et al., 1994), the Servperf model (Cronin Jr and Taylor, 1992), the retail service quality scale (Dabholkar et al., 1996) and the hierarchical and multidimensional model for service quality (Brady and Cronin Jr, 2001). However, the Servqual model has been the most cited one among researchers. It defines five service quality dimensions: tangible elements, reliability, responsiveness, empathy and assurance, in turn, composed of 22 variables (Parasuraman et al., 1994, 1988). As a matter of fact, some authors have made adaptations of the model for measuring quality specifically in hotels such as Lodgserv (Knutson et al., 1990), Holserv (Wong Ooi Mei et al., 1999), Hotelqual (Hernández Maestro et al., 2006),Resortqual (Alén González, 2004); Caltic (Moros Ochoa et al., 2016), SMSHs (Ahmad et al., 2018), Glserv (Lee and Cheng, 2018), and models without a proper name such as Servqual with 29 variables (Akbaba, 2006) and the same five dimensions (Tangibles, Reliability, Responsiveness, Assurance and Empathy) and twenty-two variables (Lestari and Laode, 2018; Lestari and Saputra, 2018; Keshavarz and Jamshidi, 2018).

As with every quality of service model, Servqual faces the task of gathering information from customers to assess their perception of the quality of service. Usually, surveys are designed and applied to retrieve this information. However, the amount of data that can be recovered with surveys can be limited by time and financial budgets. Therefore, using massive datasets from online hospitality platforms offers an exciting and valuable alternative to retrieve information, with the hindsight of being non-structured. As opposed to well-controlled surveys, information found in hospitality platforms is much more diverse because customers express their opinions with complete freedom (Zhou et al., 2020; Pourfakhimi et al., 2020). Therefore, methodologies must be devised to process and extract useful information from the data found ins online hospitality platforms. In this regard, opinion mining and machine learning methods have provided several good results in numerous contexts.

Machine learning is an area of computer science related to the understanding and design of algorithms that solve performance-measurable tasks which automatically improve through experience (Mitchell et al., 1997). The tasks intended to be solved through machine learning cover the whole complexity spectrum and reside in many fields of human knowledge. There are two branches of machine learning which are relevant for this work: supervised and unsupervised learning. In supervised learning, algorithms are expected to discover a rule set that produces an expected output given an input. The expected output is known *a priori*, as it is

usually the case that one has a training set of examples with pairs of input-output. To illustrate, consider a dataset where one has newspaper headlines as inputs and news categories as outputs. The reason for discovering a rule set that matches inputs with outputs might be to build an automatic classifier that takes any future newspaper headline and categorises it, giving it an output. Thus, the learning is said to be supervised, as one wants the algorithm to maximise the number of correctly classified headlines.

Regarding unsupervised learning, algorithms are expected to discover patterns from a collection of objects. If we stick to the same illustration, one can have a dataset with only newspaper headlines and no categories. A question that may emerge is, can we create groups of newspaper headlines without *a priori* knowledge of the actual categories? Unsupervised learning handles these sorts of tasks. In both cases, the more data there is, the better the algorithms usually perform: having more data is related to having more "experience".

In a more broad scenario, machine learning is considered an area within artificial intelligence. The impact of the latter on the hospitality industry has been more studied so far. Mainly, in recent years, researchers have devoted efforts to the following issues: understanding and measuring the possible acceptance of human-robot interaction in the hospitality sector (Lin et al., 2019; John and Cristian, 2018); producing forecasting methods to timely anticipate the demand (Önder et al., 2019); reviewing technology development of possible artificial intelligence-related products in hospitality (Chi et al., 2020), and measuring consumer experience (Sun and Norman, 2018). The last research branch has produced only a limited amount of research, mainly because in order to develop work in this area, a high interdisciplinary collaboration is needed between researchers of hospitality and tourism and researchers of mathematical methods in machine learning with the ability to gather large datasets to be analysed. Such is the case of our work.

In the quest to use the vast amount of data produced by consumers of hospitality services, machine learning and big data analytics are highlighted as promising tools to unveil hidden patterns that are important to discern consumer dynamics and study them rigorously. Machine learning has recently started to be used in this area with success. For instance, Lee et al. (2018) collected around a million hotel reviews and developed effective classification models to assess review helpfulness by using data mining techniques. Also, Martin-Fuentes et al. (2018) created a model to classify the properties offered by peer-to-peer (P2P) accommodation platforms, similar to grading scheme categories of hotels, for predicting a hotel category by taking into consideration certain quality variables. This model was applied based on information extracted from 18 million reviews from Booking.com and applied to Airbnb to predict their star rating. Ahani et al. (2019) used clustering techniques and dimensionality reduction methods to predict users interests for market segmentation purposes in Spa-hotels. Other machine learning techniques such as recurrent neural networks, long short term memory networks and convolutional networks were jointly used by Ma et al. (2018b) to extract features from text and images to study how these affect hotel review helpfulness. Even basic yet powerful methods based on word frequency analysis have been used to determine the major drivers of customers reviews (Padma and Ahn, 2020).

Moreover, Cheng and Jin (2019) used Leximancer (Smith and Humphreys, 2006) to characterise the main attributes taken into account by hotel clients when they review hotels. In the same direction, Luo and Tang (2019) used a modi-

fied version of latent aspect rating analysis to study critical features described by users when expressing their experience at a hotel. Despite the recent use of machine learning techniques in the hospitality sector, these studies do not address the quality of service measurement as their central task. However, we must emphasise that the study of comments has been proposed as an essential tool at high-level decision-making instances in other domains because it captures plenty of information about what people think or how they feel about something (e.g. (Chen et al., 2018; Agrawal et al., 2018; Luo et al., 2019; Taecharungroj and Mathayomchan, 2019; Martinez-Torres and Toral, 2019)).

Therefore, research using machine learning methods for processing online customer reviews in the hospitality sector has started to flourish recently. Still, none has focused on assessing the quality of service of the hospitality sector. Our work aims to fill this knowledge gap and to start a discussion on how to integrate robust and characterised methods such as Servqual with these vast sources of information.

3 Research Methods

In this section, we qualitatively review some of the machine learning methods used in the area of NLP, which are needed to automatically process and draw valuable information from large sets of texts. We do so by paying closer attention to the use of such NLP methods in the related work. Furthermore, we also discuss the elements from these works that are useful for our research which has the objective of qualitatively and quantitatively assess the quality of service in the hospitality sector. Here, the quality of service is understood from the point of view of Parasuraman et al. (1994) in their Servqual model, where it can be assessed by measuring the gap between the expectation and perception of service from the customers' opinions. However, we must emphasise that our approach to measuring this gap depends on the assumption that customers who post reviews on online hospitality platforms do so mainly by highlighting the experiences that either surpassed or failed to meet their expectations. In that sense, the objective of our study is to propose and test a methodology to survey the topics that concern or satisfy the customers the most and that can identify the level of satisfaction-dissatisfaction of a customer. The last point is crucial, as we are mainly interested in quantifying which aspects of the quality of service affect the most the customers' perception.

As a result of the main objective so far exposed, NLP techniques allow us to process the massive amounts of text encountered in online hospitality platforms. The backbone of the methodology that we propose in this work is familiar to many NLP applications: the text needs to be numerically represented to conduct clustering tasks or conduct prediction tasks, among other machine learning tasks. There are several alternatives to numerically represent text, grouped into feature-engineering methods and unsupervised methods. The former refers to using human and context knowledge to extract features that resemble and exhibit the semantics of text. The latter refers to language models that exploit the co-occurrence of words in different sentences to infer semantic structure in text.

An example of feature-engineering can be found in the work by Lee et al. (2018), where a model for predicting the helpfulness of reviews in TripAdvisor was built. They extracted some text features from the reviews in order to nu-

merically represent it. Some of the features are the number of characters, words, syllables and sentences, as well as some readability indices. Moreover, already implemented machine learning models (particularly Stanford's CoreNLP (Manning et al., 2014)) were used to extract more sophisticated features such as the sentiment of the reviews. Other variables, such as the reviewer age or gender, were also used. Therefore, each customer review was represented as a vector, where each component held one of the engineered features. Those vectors were subsequently used to predict each review's helpfulness, which is defined as a rate of helpful votes from other TripAdvisor users. The authors tested several classification algorithms aiming to classify a review as helpful or not. Among the used algorithms were decision trees, random forests, logistic regression and support vector machines. Consistently, random forest was the most accurate algorithm in performing the classification.

Another closely related study to predict the helpfulness of reviews from TripAdvisor and Yelp is conducted by Ma et al. (2018b). However, this is an example of unsupervised methods applied to extract latent information both in text and in user-provided photos. Later, a supervised method was trained to predict review helpfulness based on the text and photos' latent features. Mainly, latent features from the text were extracted through long-short term memory (LSTM) neural networks, which excel at discovering semantic relations between elements of a time-sequence, such as the phrases found in a customer review.

The previous studies provided different alternatives to solve a significant problem: given the vast amount of information found in online hospitality platforms, their users are faced with the impossible task of reading all of the other customers' reviews. Therefore, a way of predicting the helpfulness of customer reviews is wanted to show to the platforms' visitors only the most relevant and helpful comments that will lead that visitor to make a decision and become a customer of the platform. On the other hand, there is also a research branch that is interested in processing the same information but for a different purpose: extracting global features from the customers as an entity to deduce their main concerns and sources of satisfaction. This is the branch that our study belongs to. As might be expected, those general concepts enclosing what worries or satisfies the customers drive the qualitative assessment of the quality of service.

To illustrate, Cheng and Jin (2019) use sophisticated features by exploiting word co-occurrence matrices, which are at the core of the software Leximancer (Smith and Humphreys, 2006) that was used for text mining the reviews of Airbnb customers. The outcome of this unsupervised method revealed the central concepts used by customers to review their staying experience. Furthermore, sentiment analysis, which is focused on classifying if the sentiment of an opinion is positive or negative with respect to some topic, was performed on the reviews corpus.

Another exciting application of powerful learning techniques is given in the work of Luo and Tang (2019), where latent aspect rating analysis (LARA) was used to identify which aspects related to the staying experience had the larger impact on the rating provided by the customers from a sample of Airbnb reviews. The method is based on the popular Latent Dirichlet Allocation (LDA) model (Hoffman et al., 2010), which finds co-occurring relations between words within texts that are simultaneously assigned (probabilistically) to latent topics. These latent topics contain human-level concepts that are automatically extracted from cus-

tomer reviews. Moreover, the study also performs sentiment analysis to measure how much impact sentiment has on the review score.

All these studies are in line with the robust growth of machine learning applications in the hospitality sector that allows researchers and managers to simultaneously consider thousands of customer reviews to automatically extract the main aspects that concern or satisfy customers at specific locations. The results of these studies are closely related to the measurement of quality of service. In the research reported in this paper, we use some of the methods from these studies to build a data processing pipeline with machine learning state-of-the-art models to target the problem of extracting the main quality of service-related topics from customer reviews. Those models aim to numerically represent the semantics of reviews, which can be used to quantify rating-like levels of customer satisfaction. In what follows, we describe the design of our study as well as the dataset that was retrieved.

3.1 Data

Most of the work done in this early stage of machine learning application to customer reviews in the hospitality sector has used data from developed countries, usually including reviews in English solely. Thus, we consider it a vital research step to perform studies with data from non-English speaking countries, not only because of different language structures but also from the cultural perspective, where other sociological aspects from different countries come into play. Therefore, we decided to study two of the main touristic destinations (Ministerio de Comercio, Industria y Turismo, 2020) in a developed and a developing country: we gathered customer reviews from Madrid, Spain and Bogotá, Colombia. The online platform that we chose to gather the data was Booking.com for two reasons. First, it is a popular platform for travellers and tourists to choose and book rooms at hotels. Second, when they gather information from clients at the end of their stay, they ask for positive and negative comments, as well as rating some hospitality aspects.

Thus, the information that the customer gives to Booking.com in a review is: author's name and country, positive comment, negative comment, cleanliness score, staff score, location score, comfort score, value for money score, facilities score and free WiFi score. The scores range from 1 to 10, 1 corresponding to a bad staying experience and 10 to a pleasant experience. However, each score is not made public by Booking.com at the customer level. It is only available at the hotel level. Therefore, at the customer level, only the average of those scores is available.

We built a web crawler to gather the two cities' data, with a total count of 667 hotels in Bogotá and 1,181 in Madrid. As per the number of reviews, there were 108,563 for Bogotá and 498,361 for Madrid. It is worth noting that the comments are written in several languages, being Spanish and English the most common ones since Colombia and Spain are Spanish speaking countries. We focused on comments written only in Spanish, but the framework we propose can be applied to sets of reviews of any common language (Joshi et al., 2020).

3.2 Study design

From the literature, we encountered a handful of works performing NLP on customer reviews from online hospitality platforms, but none of them focused on the quality of service. Nonetheless, many interesting conclusions can be drawn from these works. One is that deep learning architectures perform much better than classical machine learning algorithms, especially when extracting non-supervised features from the text, i.e. semantic features, instead of engineered features. Another important finding by Luo and Tang (2019) was that a method based on LDA, a topic modelling scheme, helped study aspects and emotions present in customers reviews.

The previous elements are thus incorporated in the design of our study. We will describe the method that we propose to extract the quality of service-related features from textual information found in clients' comments on Booking.com. In general terms, we discovered the most relevant topics of positive and negative comments posted by customers of all hotels in both cities. We then created a visualisation tool that enabled us to explore the content of the topics with ease. Figure 1 shows the flow diagram of our study, which focuses on the main NLP task, namely, representing text numerically. We will explain and motivate each step of the study in the following subsections.

3.2.1 Data extraction and cleaning

First, we built a crawler using Selenium (Salunke, 2014), a tool for automating web browsing, which enabled us to rapidly download several thousands of customer reviews from all the available listings in Booking.com for Bogotá and Madrid. Second, we performed a text preprocessing, which is intended to clean the text and standardise it. The preprocessing included removing stopwords and punctuation, writing all characters in lowercase and lemmatising each token. As an example, consider the text "there were always kids playing in the Restaurant!". Stopwords do not significantly contribute to the semantics of the text, which in the example would be "there", "were", "in" and "the". After removing these stopwords, removing punctuation, and writing all characters in lowercase, the remaining words are lemmatised so that the preprocessed text is "always kid play restaurant". The lemmatisation was carried out with a Spanish lemmatiser from the open-source software spaCy.

3.2.2 Topic discovery

One of the most studied branches in NLP is topic discovery, which aims to identify groups of documents within a corpus that can be distinguished from other documents due to characteristic words that define a topic. Thus, documents that share a topic are semantically similar. For example, a topic can contain customer reviews that complain about the hotel staff kindness, and another topic can contain reviews that praise how quickly the hotel staff meets the customer's requirements.

Defining the number of topics in a corpus is an open research question in machine learning. However, the use of a topic model called Latent Dirichlet Allocation (LDA) combined with the C_V coherence to measure its quality has been
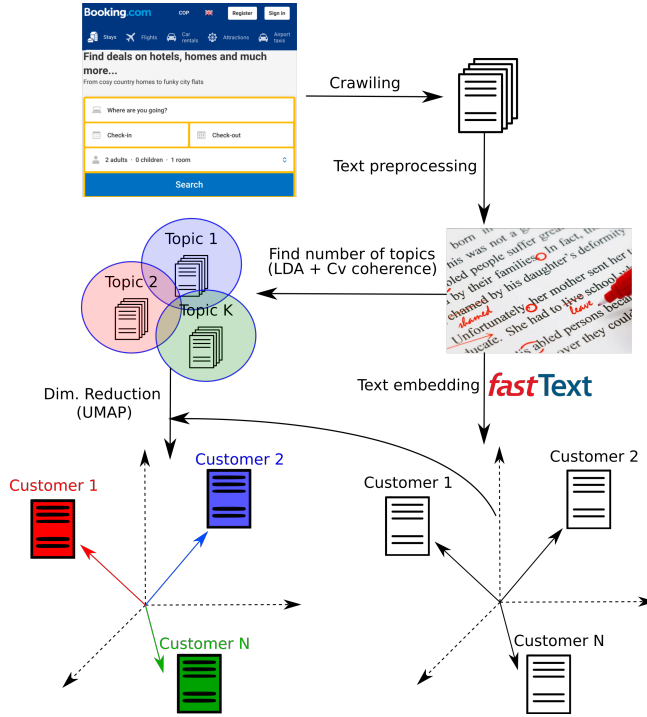
Fig. 1: Flow diagram of our work, which facilitates a qualitative assessment of the quality of service. First, customers reviews get crawled from Booking.com. Then a series of text preprocessing steps are carried out on the reviews. After that, two algorithms act on the preprocessed text: one that finds the number of quality of service-related topics in the text corpus, and another one that embeds the reviews in a real vector space. Finally, unsupervised clustering techniques are applied to the embedded vectors to group reviews in different topics at a fine-grained scale.

satisfactory (Vargas-Calderón et al., 2020; Vargas-Calderón et al., 2019) in Spanish opinions corpora. LDA (Hoffman et al., 2010) is a probabilistic algorithm that groups texts from a corpus into a number K of groups or latent topics. In other words, LDA identifies K topics in a set of texts and assigns probabilities of finding the contents of each topic within each text. To measure the quality or coherence of the topics, i.e. how well-defined a topic is, we use the C_V coherence, a measure of the correlation of top occurrent words in a topic (see appendix B for details). As an example, if we tell LDA to find five topics in a set of customer reviews, we might find that one topic contains customer reviews saying that the hotel staff never cleans the bathroom, others saying that the towels were of inferior quality. Indeed, those are bathroom-related reviews, but if we had told LDA to find six topics, this topic is likely split into two: one related to the cleanliness of the bathroom and another one related to the quality of the towels. In this case, the C_V coherence is greater than the case where only five topics were used. However, if the number of topics is too high, the C_V coherence decreases because there exists a high similarity between some topics. Indeed, the C_V coherence has shown outstanding

agreement with human judgements of the interpretability of the topics extracted by LDA (Röder et al., 2015; Syed and Spruit, 2017). An important remark is that LDA gives the probability that each text belongs to a topic, which allows a text to contain elements from different topics. For instance, a customer may complain about the quality of the bed but can also praise the hotel restaurant chef, which is reflected on the LDA model by assigning to this review a mixed probability of belonging to two topics.

3.2.3 Visualisation of reviews and topics

An intuitive and accurate way of exploring an extensive reviews dataset is to navigate through representative reviews of each topic. Here we describe the inner working of a visualisation tool that we designed to explore and read few but representative reviews for every topic that LDA and C_V coherence can discover.

A helpful way of visualising reviews and topics is to consider their vector representations, use a dimensionality reduction algorithm (which takes vectors from a high-dimensional space to a 2-dimensional space), and plot each review as a coloured dot, where the colour indicates the topic. In our case, we use an algorithm called uniform manifold approximation and projection (UMAP) (McInnes et al., 2018). What UMAP does is to learn topological features of N-component vectors to find accurate projections onto a lower-dimensional vector space that tries to preserve the distance from the high-dimensional space to the low-dimensional space. More formally stated, UMAP retains the original vectors' manifold structure in the high-dimensional vector space and projects it onto a low-dimensional vector space.

The resulting plot from reducing the dimension of vectors with UMAP will show a map of the corpus, where points near each other mean that their respective documents are semantically similar. When using the vector representation of LDA, documents from different topics appear somewhat mixed in these plots (not shown) because the vector representation is done in a K-dimensional vector space, where K is the number of topics, which in the case of this work is of the order of 10.

Therefore, to improve this, we are motivated to use vector representations of the reviews containing more fine-grained semantic information. An excellent algorithm to do this is the celebrated FastText text vector embedding model (Bojanowski et al., 2016; Joulin et al., 2016), which efficiently uses the co-occurrence of words in the text in order to assign vectors of N components to a piece of text. The vectors store in their components semantic features and linguistic contexts from the text so that two similar texts will have assigned to them two similar vectors. In the lower-left part of fig. 1, this is depicted in a 3D vector space (so that $N = 3$) with three example customer reviews, but in general, one can choose the dimension N of the vector space where the embedded vectors are. The readers interested in the underlying mechanism of text embeddings are referred to appendix A where we give essential concepts of FastText, or to the FastText papers (Bojanowski et al., 2016; Joulin et al., 2016) as well as the Word2Vec paper (Mikolov et al., 2013), which give much more mathematical detail about this method. This vector representation achieved by FastText is much richer than LDA's.

Therefore, the reviews and topics' visualisation can be achieved by computing vector representations of the reviews in an N-dimensional vector space using Fast-Text ($N = 100$ for this work). Then, these vectors are reduced to 2-dimensional

vectors using UMAP. Next, a scatter plot is generated, where the colour of each point is determined by how LDA assigns the corresponding review to a topic.

The only missing detail is that only representative reviews should be plotted, as there might be reviews containing several topics, thus making it challenging to explore specific topics. To filter out only representative reviews for each topic, we choose the reviews that belong to that topic with a probability higher than a selected threshold. In our case, we chose this threshold to be 0.8, although the selection of this value is entirely arbitrary and depends on how many points the user would like to see plotted. It is worth noting that in the topic exploration phase, we ignore those reviews whose maximum probability of being assigned to a topic is less than 0.8; however, the topic probability distribution for each review can be used, for instance, to assess the magnitude of each positive or negative topic within a hotel by computing $M_T = \sum_{r \in R_H} P(r \in T)$, where M_T denotes this magnitude for a topic T, R_H is the set of reviews of hotel H, and $P(r \in T)$ is the probability that the review r belongs to topic T. The analysis of these magnitudes for each topic in a hotel, compared with other hotels, might give a qualitative and quantitative market comparison.

All of this process is depicted in fig. 1, and is done in 4 sets of reviews: positive reviews from customers in Bogotá, negative reviews from customers in Bogotá, positive reviews from customers in Madrid and negative reviews from customers in Madrid. Our framework allows us to find topics and visualise reviews in those topics for each of the mentioned groups. Moreover, this framework can be applied to any group of customer reviews: one can consider customer reviews from a whole country, for a single hotel or a group of hotels, etc.

3.2.4 Scores and topics

Finally, we will show the score distributions for each topic found in the negative and positive comments in Bogotá and Madrid. Each review is composed of a positive, a negative comment and a score. Therefore, positive topics are not necessarily related to high scores, nor negative topics are necessarily related to low scores. In fact, these score distributions will show us which positive (or negative) topics are essential for the customer to assign a high or low score. For example, a customer might have a terrible experience but might have liked the hotel restaurant's food. Therefore, this customer might give a negative comment, reviewing which aspects he or she did not like about their stay, and also might give a positive comment saying that the food at the restaurant was good; however, since the staying experience was mostly bad, the customer assigns a low score. This means that even though the food was good, it is not a determinant factor to compensate for the other hotel features that the customer perceived as bad. We must emphasise that the perception of a hotel as better, when compared to another, is not driven by standardised service or facilities, but rather it is driven by the average gap between the customers' expectations and perceptions of the quality of service (this work proves this claim).

4 Results and Discussion

We will interpret the results provided by our framework by performing constant comparisons to the widely used quality of service model Servqual since it is one of the most researched and robust quality of service models in the hospitality industry. We will see that our framework allows us to recognise some of the Servqual model's dimensions but extends it and limits it to the topics that are more relevant to the customers.

The first step of our method aims to estimate the number of topics in each of four customer reviews sets: positive and reviews from Bogotá and Madrid. Figure 2 shows the measurement of the C_V coherence as a function of the number of topics, where the coherence reaches a maximum for 12, 12, 7 and 9 topics for positive reviews from Bogotá, positive reviews from Madrid, negative reviews from Bogotá and negative reviews for Madrid, respectively.
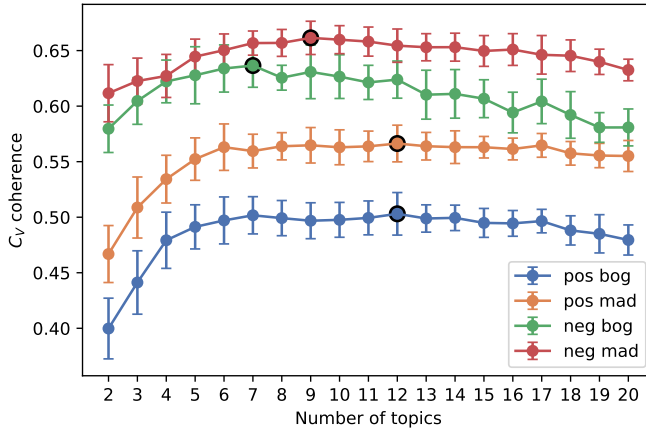

Fig. 2: C_V coherence as a function of the number of topics in an LDA model. Since LDA is probabilistic, we measured the coherence by independently training 20 times an LDA model for each number of topics and each set of comments (we found 20 times to be a good number, as adding more training runs did not affect coherence variance). The data points are the average of those measurements, and the error bars are their standard deviation. Circled in black are the maximum coherence for each set of reviews.

After obtaining the FastText vector for each review and performing dimensionality reduction with UMAP, we ended up with visualisations that allowed the exploration of representative reviews for each of the identified topics. The visualisations of positive topics from Bogotá and Madrid are shown in fig. 3, where each point in the figures represents a single customer review, representative of a latent topic. For Bogotá, 9.5% of the positive reviews were representative (i.e. the highest probability to be assigned to a topic surpassed 80%), and for Madrid, 6.4% of the positive reviews were representative. The colour of the points identifies different

latent topics. Points close to each other have a similar FastText vector representation, which is why each figure contains several well-packed clusters, separated from the rest of the comments. This visual representation allows a better human interpretation of the topics that are important for hotel clients. The clusters are easily identifiable and have well-defined topics, which were annotated and shown in the figures.

We see that there are many similarities in the topics found in positive reviews from Bogotá (top panel of fig. 3) and Madrid (bottom panel of fig. 3), but we also see some differences. In general, reviews provide information that is not usually considered by quality of service models in the hospitality industry, such as adaptations of the widely used Servqual model. Nonetheless, we emphasise that all of this information is contained within the general dimensions defined by the Servqual model. Most of the positive reviews refer to variables included in the tangible elements Servqual dimension, meaning that having comfortable bedrooms, clean spaces, beautiful room views and good breakfasts are crucial for the clients when giving a positive review. Also, the kindness of the staff is vital in positive reviews for both cities. However, there are other elements of great value for the clients which are not usually explicitly included in the Servqual model, such as good location, airport transfer services, closeness to metro stations, closeness to the centre (in Madrid), closeness to USA Embassy or airport (in Bogotá). These topics, however, can also be associated to the tangible elements Servqual dimension. Also, the booking system seems to be meaningful and valued in Bogotá, and we suppose that clients in Madrid assume an excellent booking system as given.

These examples show that our method automatically shows the main topics on which hotels should focus their attention to induce in their clients a positive perception of their stay and can do so at any granularity level (in this case, at a city level). This is further taken advantage of when examining the negative reviews, as shown in fig. 4, where 14.3% of the negative reviews were representative for Bogotá 7.5% for Madrid.

In the case of negative reviews, most topics also refer to Servqual's tangible elements dimension. Some of these topics are temperature in the room, poor maintenance in the installations, small room/bathroom, noise, and lousy WiFi. In Bogotá (top panel of fig. 4), clients complain about uncomfortable bed/pillows, no credit card payment option, little variety in breakfast or a limited breakfast schedule.

On the other hand, in Madrid (bottom panel of fig. 4), topics such as room facing the inner courtyard or lack of an elevator and lack of bellboy stand out. It is observed that negative reviews stress cultural differences. Bogotá contains many negative reviews related to noise, which is a factor that must be taken care of by Bogotá's hospitality sector by sound-proofing hotel rooms. As the food is relatively cheap in Colombia, no price complaints are made in Bogotá, but they highlight in Madrid. Also, the option of paying with a credit card is not a central topic in Madrid, but it is in Bogotá. This can be due to a cultural aspect of Colombia, where many local hotels do not support credit card payment[1].

---

[1] As a matter of fact, in the past two years, Colombian government is applying a public policy where all kind of businesses have to have electronic payment options for electronic billing (see Ruling 42 of 2020 from Colombia's DIAN).
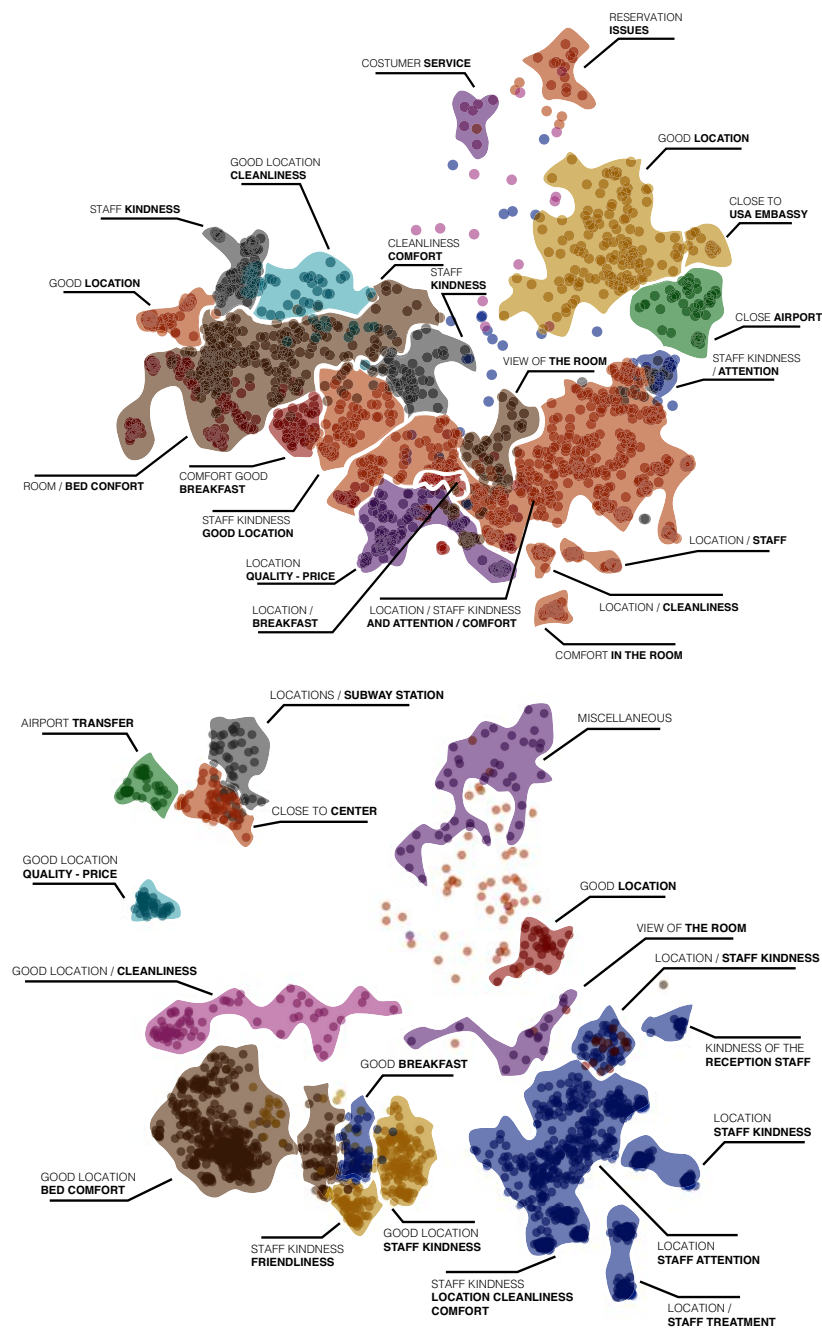
Fig. 3: 2D projection found by UMAP of clusters coloured by LDA analysis of the FastText representation of positive reviews from Bogotá (top panel) and Madrid (bottom panel). Each point in the figure represents a single customer review.
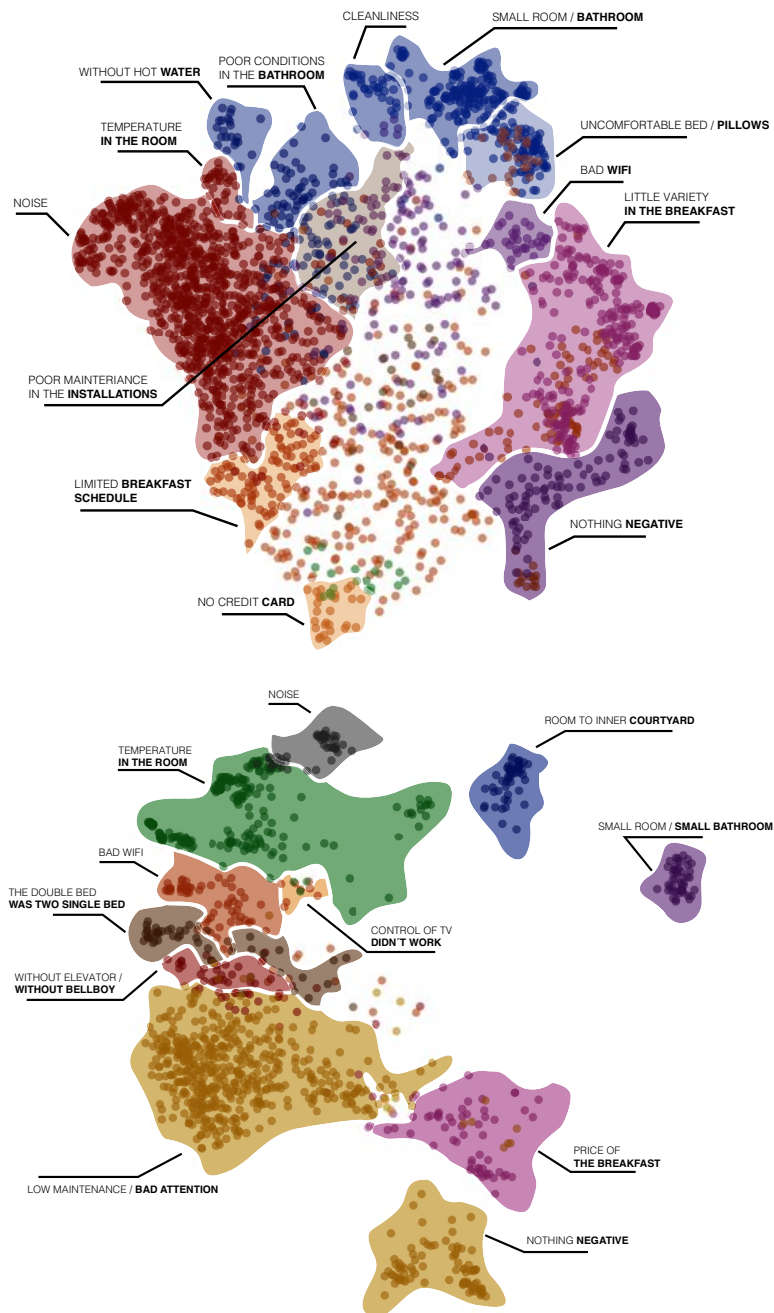
Fig. 4: Same as fig. 3 but for negative reviews from Bogotá (top panel) and Madrid (bottom panel).

To help the reading of figs. 3 and 4, table 1 lists the main topics found in the positive and negative reviews of both cities, along with their associated Servqual dimension. The association of Servqual dimensions to each major topic found through our framework is done by examining the relation between relevant comments for each topic and the definition of each Servqual dimension. As explained by Zeithaml et al. (2018), there are five Servqual dimensions:

— *Tangible elements.* These are related to physical appearance, physical facilities, infrastructure, equipment, materials and staff.
— *Reliability.* It refers to fulfilling the promised service to the client carefully and reliably. This implies that the company is faithful to every feature of the service they say they will provide.
— *Responsiveness.* This refers to being available to help clients and to provide adequate and quick assistance in any need (question, complaint, problem) related to the service being provided.
— *Assurance.* This is the capacity to inspire credibility and trust in the client based on the employee's knowledge, skill and attention.
— *Empathy.* It refers to the custom attention offered by the company to its clients. It must be transmitted through a custom service, adapted to the clients' needs.

These definitions allow the categorisation of fine-grained topics found in figs. 3 and 4 into the Servqual dimensions. However, the fine-grained topics are only aspects of more prominent topics, which were automatically identified with our framework, and shown in table 1. In what follows, we will provide an association between major topics and Servqual's dimensions.

For positive reviews, we see that the following topics are related to tangible elements: bed comfort, large bedroom, cleanliness, good breakfast, and view of the room. The airport transfer-related topic can be classified into the empathy Servqual dimension if the hotel presents a service to their customers to ease airport transfers; but, if the hotel possesses buses or cars to provide such a service, this topic can also be related to the dimension of the tangible elements. Reservation issues and their solutions, as well as the miscellaneous topic (this topic includes aspects related to problems about facilities that were effectively solved by the hotel staff), are related to the responsiveness Servqual dimension. Staff kindness or attention and quality-price relation are related to reliability. Customer assistance is related to empathy. Location-related topics is a significant finding of our study: customers value how near the hotel is from certain points of interest within the city or how easy it is to get from the hotel to those points of interest. We should mention that, even though Booking.com asks customers to evaluate the hotel location, our framework identifies this topic without *a priori* knowledge about this request. These aspects are not usually related in other studies to tangible elements, but we identify that they can be included in this Servqual dimension. However, it is important to remark that tangible elements use to be linked to features inside the hotel, whereas location is related to the hotel position within the city relative to points of interest. Thus, location-related topics become also an important feature for the hotel's value proposal, especially when promoting the hotel, for instance, through the Booking.com website. Moreover, some location topics that are related to security perception (in the criminal sense), could also be associated to the assurance Servqual dimension.

Regarding negative reviews, room and bathroom issues are related to tangible elements and reliability, as sometimes hotels do not provide the same offered services. Even though some of these issues can be directly associated with the tangible elements Servqual's dimension, there are specific aspects captured by our framework that do not tend to be in customer service surveys, such as the availability of hot water or the bathroom size. Also, in the topic of room and bathroom issues, there is quite a common reliability-related issue that stands out in the aspects retrieved by our framework, which is that hotels mistakenly provide two individual beds when one double bed was booked. Noise (absence of sound-proof environments), lousy WiFi, no elevator, poor maintenance in the installations, payment methods (e.g. the unavailability of credit card payments), and breakfast issues can also be classified in the tangible elements dimension. Billing issues are related to two Servqual categories: reliability and assurance. Commonly, clients express billing issues because there were errors in the billing, which increase distrust in the hotel staff. Regarding empathy, it encloses topics such as bad attention. An interesting topic is dangerous location, which is usually not related directly to any variable in the several adaptations of the Servqual model to the hospitality industry. Reviews related to this topic show that the hotel is located in places perceived as insecure. Therefore, the dangerous location topic is more related to the hotel's environment, falling into the tangible elements and also assurance Servqual's dimensions. Hotel management should be aware of this situation to improve the safety perception of their clients.

Moreover, table 1 presents the percentage of reviews related to each topic, as well as the salience of each topic, which we compute by summing the probabilities that a word belongs to a topic considering the top 10 words for each topic. This table shows one of our main results: there are important topics for the customers that are usually not found in the surveys from different adaptations of the Servqual model, such as the location, with a significant percentage of positive reviews: 34.8% for Bogotá and 21.8% for Madrid. Also, table 1 and figs. 3 and 4 show that some service elements are not equally valued in different countries because distinct tourist populations have diverse expectations. An example of this is the clear presence of topics related to billing issues or payment methods in Bogotá and not in Madrid. Another notable point is the large amount of concern displayed in the reviews regarding room/bathroom issues such as temperature in the room (in Madrid), no hot water (in Bogotá), room/bathroom size, poor bathroom conditions, wrong bed size, among others. Indeed, this is a neuralgic point, especially in Madrid, taking almost half of the negative reviews.

On the other hand, from a linguistic point of view, each topic's salience expresses how unique the vocabulary used to describe positive or negative issues is. For instance, very high saliences are obtained for bed comfort and quality-price relationship in positive reviews concerning other topics. Regarding negative reviews, the largest saliences are related to room/bathroom and breakfast issues and noise (in Madrid).

Finally, we study how these identified topics are related to the scores given by the customers. Notably, we answer how frequently are positive and negative topics related to different ranges of scores. To do this, in fig. 5 we show the score distribution (through box-plots) of the positive and negative relevant reviews identified for each positive or negative topic in Bogotá. In both types of reviews, we found statistically significant differences between the scores using the ANOVA test ($p \ll 0.05$).

Table 1: Main latent topics depicted in figs. 3 and 4 for positive and negative reviews and their associated Servqual dimensions. The first column presents the main major topics for positive and negative reviews. The second column relates Servqual dimensions to those topics. Rows are grouped with green/white rows based on those Servqual dimensions. The last two columns present numbers of the form XX/YY, where XX stands for the probability percentage assigned to a topic by the LDA model taking into account all documents, and YY stands for the salience of each topic, by summing the probability of the top 10 words assigned to each topic by the LDA model.

		City	
Positive reviews	Servqual dimension	Bogotá	Madrid
Bed comfort	Tangible elements	12.6/0.63	13.8/0.69
Large bedroom			7.3/0.24
Cleanliness		8.6/0.43	8.1/0.47
Good breakfast		4.3/0.41	6.8/0.32
Location-related		34.8/0.47	21.8/0.35
View of the room		8.5/0.45	
Airport Transfer	Tangible elements or Empathy		7.3/0.22
Reservation issues	Responsiveness	5.5/0.30	
Miscellaneous			4.9/0.27
Quality-price	Reliability	6.1/0.48	5.0/0.71
Staff kindness/attention		12.9/0.42	25.0/0.71
Customer assistance	Empathy	6.9/0.24	
Negative reviews			
Room/bathroom issues	Tangible elements and reliability	19.8/0.30	44.9/0.30
Noise	Tangible elements	20.1/0.23	9.8/0.37
Bad WiFi			9.0/0.18
No Elevator			7.9/0.32
Poor maintenance in the installations		11.0/0.18	
Payment methods		11.3/0.15	
Breakfast issues		14.2/0.33	14.4/0.20
Billing issues	Reliability and assurance	15.2/0.18	
Bad attention	Empathy		14.0/0.17
Dangerous location	Tangible elements and assurance	8.4/0.16	

Furthermore, we performed the Tukey's honestly significantly differenced (HSD) test (Tukey, 1949) on the different pairs of topics for positive and negative reviews, finding almost all of the differences statistically significant ($p < 0.05$). We see that a group of positive reviews (recall that customers are asked to write a positive and a negative comment) related to reservation issues and how they were solved have a low score distribution with respect to other positive topics. Also, topics related to staff (customer service, staff attention and staff kindness) offer a fascinating insight: while customer service and staff attention get mentioned in positive reviews, their score distribution is much lower than the staff kindness topic score distribution. We hypothesise that customers expect some standard customer service and staff attention, but they really value kindness, which is why this topic is correlated with very high scores. Regarding negative reviews, there is a small topic (meaning

that only 8.4% of reviews consider this topic, not highlighted in fig. 4) with reviews about the surroundings of the hotel being dangerous (in a security and criminal sense) that has a high score distribution. This means that people do not penalise with low scores when the hotel is situated at a dangerous-looking location but give more importance to other positive aspects of the hotel. On the other hand, topics with a larger percentage of relevant reviews, such as having small rooms or bathrooms, billing issues and noisy rooms, obtain lower scores. In Bogotá, these topics are the most common concerns among customers and are correlated with low scores (below 6/10).
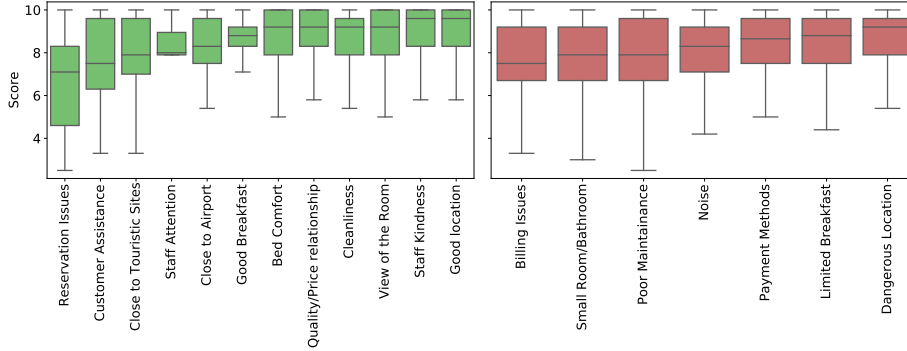

Fig. 5: Score distribution of relevant reviews with positive (left) and negative (right) topics for hotels in Bogotá. As before, relevant reviews are selected as those whose probability of belonging to a topic is above a threshold (80%).

Regarding Madrid, differences in scores through ANOVA and Tukey's HSD test were also statistically significant ($p \ll 0.05$) between topics. As in Bogotá, good location and staff kindness/attention are important positive topics correlated with high score distributions. Interestingly, many customers point out that rooms are spacious, but this is a feature that is not necessarily correlated with high scores. With respect to negative reviews, small bedrooms or bathrooms, wrong bed sizes (e.g. customers book a double bed but get two single beds instead), and bad attention from the staff are main topics of concern for customers correlated with low score probabilities. On the other hand, sometimes hotels do not include breakfast by default, and many customers find this disappointing. Thus, they highlight this issue as a negative topic when they give low scores (below 6/10). Also, there is a high frequency of negative reviews associated with street noise for low scores, meaning that it is an essential source of discomfort for customers.

5 Conclusions

In this work, we presented a framework to automatically extract the quality of service-related features from large databases of hotel customer reviews. We exemplified the use of our framework with customer reviews from Bogotá, Colombia and Madrid, Spain, in Booking.com. The framework is based on machine learning
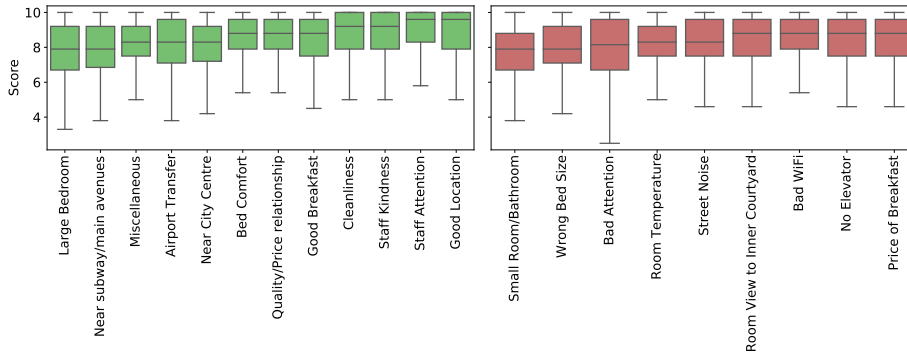
Fig. 6: The same as fig. 5 but for Madrid.

algorithms for latent topic discovery, dimensionality reduction and vector representation of text. By combining these algorithms, we generated visualisations of big data sets of comments by clustering them and showing the most representative ones in each cluster. The results obtained show that the visualisations aid human readability of many comments and show essential information, which could aid decision-making in the hospitality sector. Moreover, we were able to discover the critical aspects for clients when reviewing positively or negatively hotels in the studied cities.

5.1 Research implications

Our research's theoretical implications aim to discuss the validity of using large online datasets from online hospitality platforms when assessing the quality of service through non-structured data. Quality of service models have been characterised by a static set of aspects based on customers experiences that are measured and gathered by researchers. The availability of large datasets in online hospitality platforms is changing and will change the methodologies for collecting information and for designing quality of service models. Moreover, the power of machine learning algorithms is starting to be used as a means of building quality of service models that are easily adaptable to the dynamics of the hospitality sector. This work contributes to this new but promising area of research.

Also, in our analyses, we encountered that cultural differences in segments of the global population produced different expectations of their hotel stay. An interesting observation is that Colombian clients rated (from 1 to 10 in Booking.com) Madrid hotels about 5% higher than the average Spanish hotel client, whereas Spanish clients rated Bogotá hotels about 5% lower than the average Colombian hotel client. This can also be seen in the comments, as there are different main topics in the two cities. For instance, many people complain about noise in Bogotá (probably because not all hotels have soundproof rooms), and in Madrid, they do not as much. Another example is that clients value being close to metro stations in Madrid because it is an efficient and comfortable public transportation system. On the other hand, in Bogotá, the public transportation system is not efficient nor comfortable and does not densely connect many parts of the city.

Our framework also has tremendous practical implications. It directly enables managers and researchers to identify the critical quality of service topics that affect (positively or negatively) the perceived quality of service, thus providing essential information to propose and implement improvement strategies. In fact, we showed that some of these topics are not usually taken into account by well-established quality of service models such as adaptations of the Servqual model to the hospitality industry.

Moreover, the topics discovered and explored through our framework can be compared with customer ratings. Although ratings are not as telling and expressive as customer reviews, they give a rough estimate of the customers' level of satisfaction. Therefore, from the automatically identified topics, one can establish a connection between some topics and rating distributions, as was also shown in this work.

5.2 Limitations and future work

Although our research used a large dataset, it only explored two capital cities, one online hospitality platform and one language: Spanish. Our framework has the capability to escalate to many other destinations and other platforms such as Airbnb and other P2P and business-to-consumer (B2C) platforms. It would be important to analyse the relevance of including other languages in future research to identify not only cultural differences but also social, psychological and personal differences that influence the perception of service quality of hotel customers in different cities.

Since the amount of text available from customer reviews influences the robustness of machine learning methods, it is important to note that our method could be limited to scenarios in which hotels do not have enough customer reviews. This aspect will be addressed in future work to analyse the effectiveness of our method in small/not popular/low-class hotels with few reviews with respect to high/popular/famous/high-class hotels with many reviews. In this scenario, we remark that doing a stratified analysis by hotel category might reveal topics that are important for certain clientele, but not for others.

Future work includes gathering customer reviews from other cities to achieve better latent topic models. More importantly, we aim for a pilot implementation of this framework as a quality of service assessment tool in a hotel willing to execute plans to improve the quality of service perception by analysing the information provided by our framework.

Finally, a crucial research avenue is a theoretical and practical study of integrating large online hospitality platforms as information sources for robust quality of service models such as Servqual. We foresee two main research lines: how to automatise retrieval information to score the variables of the Servqual model from online reviews; and how to adapt the Servqual model to specific locations and sub-markets, where customers worry about some specific service features.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abubakar AM, Ilkan M, Al-Tal RM, Eluwole KK (2017) ewom, revisit intention, destination trust and gender. Journal of Hospitality and Tourism Management 31:220–227

Agrawal V, Bhakar S, Rana PS, Tiwari D (2018) Prediction of online perceived service quality using spider monkey optimisation. World Review of Science, Technology and Sustainable Development 14(4):376–393

Ahani A, Nilashi M, Ibrahim O, Sanzogni L, Weaven S (2019) Market segmentation and travel choice prediction in spa hotels through tripadvisor's online reviews. International Journal of Hospitality Management 80:52 – 77, DOI https://doi.org/10.1016/j.ijhm.2019.01.003

Ahmad SZ, Ahmad N, Papastathopoulos A (2018) Measuring service quality and customer satisfaction of the small-and medium-sized hotels (smshs) industry: lessons from united arab emirates (uae). Tourism Review

Akbaba A (2006) Measuring service quality in the hotel industry: A study in a business hotel in turkey. International journal of hospitality management 25(2):170–192

Alén González ME (2004) Evaluación de la calidad percibida por los clientes de establecimientos termales a través del anélisis de sus expectativas y percepciones. Revista galega de economía 13(1-2):5–22

Anderson EW, Sullivan MW (1993) The antecedents and consequences of customer satisfaction for firms. Marketing Science 12(2):125–143, DOI 10.1287/mksc.12.2.125, https://doi.org/10.1287/mksc.12.2.125

Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information. arXiv preprint arXiv:160704606

Brady MK, Cronin Jr JJ (2001) Some new thoughts on conceptualizing perceived service quality: a hierarchical approach. Journal of marketing 65(3):34–49

Buhalis D (2019) Technology in tourism-from information communication technologies to etourism and smart tourism towards ambient intelligence tourism: a perspective article. Tourism Review

Buhalis D, Harwood T, Bogicevic V, Viglia G, Beldona S, Hofacker C (2019) Technological disruptions in services: lessons from tourism and hospitality. Journal of Service Management

Chen Y, Wang J, Lai G (2018) Research on improving the government service quality by public comments monitoring: Take suburb park an example. In: 2018 15th International Conference on Service Systems and Service Management (ICSSSM), IEEE, pp 1–5

Cheng M, Jin X (2019) What do airbnb users care about? an analysis of online review comments. International Journal of Hospitality Management 76:58 – 70, DOI https://doi.org/10.1016/j.ijhm.2018.04.004

Chi OH, Denton G, Gursoy D (2020) Artificially intelligent device use in service delivery: a systematic review, synthesis, and research agenda. Journal of Hospitality Marketing & Management 0(0):1–30, DOI 10.1080/19368623.2020.1721394, URL https://doi.org/10.1080/19368623.2020.1721394, https://doi.org/10.1080/19368623.2020.1721394

Cronin Jr JJ, Taylor SA (1992) Measuring service quality: a reexamination and extension. Journal of marketing 56(3):55–68

Dabholkar PA, Thorpe DI, Rentz JO (1996) A measure of service quality for retail stores: scale development and validation. Journal of the Academy of marketing Science 24(1):3

Dhar RL (2015) Service quality and the training of employees: The mediating role of organizational commitment. Tourism Management 46:419 – 430, DOI https://doi.org/10.1016/j.tourman.2014.08.001

Douven I, Meijs W (2007) Measuring coherence. Synthese 156(3):405–425

Ghotbabadi AR, Baharun R, Feiz S (2012) A review of service quality models. In: 2nd International Conference on Management, pp 1–8

Gronroos C (1984) A service quality model and its marketing implications. European Journal of marketing

Harris ZS (1954) Distributional structure. Word 10(2-3):146–162

Hernández Maestro RM, Muñoz Gallego PA, Santos Requejo L (2006) Calidad objetiva y su relación con la formación y la satisfacción del empresario: El caso de los alojamientos rurales españoles. Universidad de Salamanca (España) Facultad de Economía y Empresa

Hoffman MD, Blei DM, Bach F (2010) Online learning for latent dirichlet allocation. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, Curran Associates Inc., USA, NIPS'10, pp 856–864

Instituto Nacional de Estadística (2020) Un retrato de nuestros turistas. URL https://www.ine.es/ss/Satellite?L=es_ES&c=INECifrasINE_C&cid=1259952806229&p=1254735116567&pagename=ProductosYServicios%2FINECifrasINE_C%2FPYSDetalleCifrasINE

John B, Cristian M (2018) Beware hospitality industry: the robots are coming. Worldwide Hospitality and Tourism Themes 10(6):726–733, DOI 10.1108/WHATT-07-2018-0045, URL https://doi.org/10.1108/WHATT-07-2018-0045

Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M (2020) The state and fate of linguistic diversity and inclusion in the nlp world. arXiv preprint arXiv:200409095

Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv preprint arXiv:160701759

Keshavarz Y, Jamshidi D (2018) Service quality evaluation and the mediating role of perceived value and customer satisfaction in customer loyalty. International Journal of Tourism Cities 4(2):220–244

Kim S, Kandampully J, Bilgihan A (2018) The influence of ewom communications: An application of online social network framework. Computers in Human Behavior 80:243–254

Knutson B, Stevens P, Wullaert C, Patton M, Yokoyama F (1990) Lodgserv: A service quality index for the lodging industry. Hospitality Research Journal 14(2):277–284

Lai IK, Hitchcock M, Yang T, Lu TW (2018) Literature review on service quality in hospitality and tourism (1984-2014). International Journal of Contemporary Hospitality Management

Lamest M, Brady M (2019) Data-focused managerial challenges within the hotel sector. Tourism Review

Lee PJ, Hu YH, Lu KT (2018) Assessing the helpfulness of online hotel reviews: A classification-based approach. Telematics and Informatics 35(2):436–445

Lee WH, Cheng CC (2018) Less is more: A new insight for measuring service quality of green hotels. International Journal of Hospitality Management 68:32–40

Lestari YD, Laode M (2018) Service innovation of 3/2 star hotel in bandung. The Journal of Asian Finance, Economics and Business (JAFEB) 5(3):73–80

Lestari YD, Saputra D (2018) Market study on hospitality sector: Evidence from 4/5-star hotel in bandung city indonesia. International Journal of Business & Society 19(1)

Lin H, Chi OH, Gursoy D (2019) Antecedents of customers' acceptance of artificially intelligent robotic device use in hospitality services. Journal of Hospitality Marketing & Management 0(0):1–20, DOI 10.1080/19368623.2020.1685053, URL https://doi.org/10.1080/19368623.2020.1685053, https://doi.org/10.1080/19368623.2020.1685053

Lin HC, Han X, Lyu T, Ho WH, Xu Y, Hsieh TC, Zhu L, Zhang L (2020) Task-technology fit analysis of social media use for marketing in the tourism and hospitality industry: a systematic literature review. International Journal of Contemporary Hospitality Management

Luo Q, Chen Y, Chen L, Luo X, Xia H, Zhang Y, Chen L (2019) Research on situation awareness of airport operation based on petri nets. IEEE Access 7:25438–25451

Luo Y, Tang RL (2019) Understanding hidden dimensions in textual reviews on airbnb: An application of modified latent aspect rating analysis (lara). International Journal of Hospitality Management 80:144 – 154, DOI https://doi.org/10.1016/j.ijhm.2019.02.008

Ma E, Cheng M, Hsiao A (2018a) Sentiment analysis–a review and agenda for future research in hospitality contexts. International Journal of Contemporary Hospitality Management

Ma Y, Xiang Z, Du Q, Fan W (2018b) Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. International Journal of Hospitality Management 71:120 – 131, DOI https://doi.org/10.1016/j.ijhm.2017.12.008

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp 55–60

Mariani M (2019) Big data and analytics in tourism and hospitality: a perspective article. Tourism Review

Mariani M, Baggio R, Fuchs M, Höepken W (2018) Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management

Martin-Fuentes E, Fernandez C, Mateu C, Marine-Roig E (2018) Modelling a grading scheme for peer-to-peer accommodation: Stars for airbnb. International Journal of Hospitality Management 69:75–83

Martinez-Torres M, Toral S (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, DOI https://doi.org/10.1016/j.tourman.2019.06.003

McInnes L, Healy J, Melville J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints 1802.03426

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Ministerio de Comercio, Industria y Turismo (2020) Centro de información turística de colombia. URL `http://www.citur.gov.co/`

Mitchell TM, et al. (1997) Machine learning

Moro S, Esmerado J, Ramos P, Alturas B (2019) Evaluating a guest satisfaction model through data mining. International Journal of Contemporary Hospitality Management

Moros Ochoa M, Vázquez JCR, Nieto GYC, Viloria A, Ariza-Salazar J (2016) Adaptation of the "caltic" service quality model in the tourism sector. International Journal of Control Theory and Applications ISSN pp 0974–5572

Önder I, Gunter U, Scharl A (2019) Forecasting tourist arrivals with the help of web sentiment: A mixed-frequency modeling approach for big data. Tourism Analysis 24(4):437–452

Padma P, Ahn J (2020) Guest satisfaction & dissatisfaction in luxury hotels: An application of big data. International Journal of Hospitality Management 84:102318, DOI https://doi.org/10.1016/j.ijhm.2019.102318, URL `http://www.sciencedirect.com/science/article/pii/S0278431919301549`

Parasuraman A, Zeithaml VA, Berry LL (1988) Servqual: A multiple-item scale for measuring consumer perceptions of service quality. Journal of retailing 64(1):12

Parasuraman A, Zeithaml VA, Berry LL (1994) Reassessment of expectations as a comparison standard in measuring service quality: implications for further research. Journal of marketing 58(1):111–124

Pourfakhimi S, Duncan T, Coetzee WJ (2020) Electronic word of mouth in tourism and hospitality consumer behaviour: state of the art. Tourism Review

Rahmani K, Gnoth J, Mather D (2018) Tourists' participation on web 2.0: A corpus linguistic analysis of experiences. Journal of Travel Research 57(8):1108–1120

Rahmani K, Gnoth J, Mather D (2019) A psycholinguistic view of tourists' emotional experiences. Journal of Travel Research 58(2):192–206

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, WSDM '15, pp 399–408, DOI 10.1145/2684822.2685324

Salunke SS (2014) Selenium Webdriver in Python: Learn with Examples, 1st edn. CreateSpace Independent Publishing Platform, North Charleston, SC, USA

Septianto F, Chiew TM (2018) The effects of different, discrete positive emotions on electronic word-of-mouth. Journal of Retailing and Consumer Services 44:1–10

Smith AE, Humphreys MS (2006) Evaluation of unsupervised semantic mapping of natural language with leximancer concept mapping. Behavior Research Methods 38(2):262–279, DOI 10.3758/BF03192778

Sun TVW, Norman A (2018) Exploring customer experiences with robotics in hospitality. International Journal of Contemporary Hospitality Management 30(7):2680–2697, DOI 10.1108/IJCHM-06-2017-0322, URL `https://doi.org/10.1108/IJCHM-06-2017-0322`

Syed S, Spruit M (2017) Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp 165–174, DOI 10.1109/DSAA.

2017.61

Taecharungroj V, Mathayomchan B (2019) Analysing tripadvisor reviews of tourist attractions in phuket, thailand. Tourism Management 75:550 – 568, DOI https://doi.org/10.1016/j.tourman.2019.06.020

Tukey JW (1949) Comparing individual means in the analysis of variance. Biometrics pp 99–114

Vallejo JM, Redondo YP, Acerete AU (2015) Las características del boca-oído electrónico y su influencia en la intención de recompra online. Revista Europea de Dirección y Economía de la Empresa 24(2):61 – 75, DOI https://doi.org/10.1016/j.redee.2015.03.002

Vargas-Calderón V, Dominguez MS, Parra-A N, Vinck-Posada H, Camargo JE (2020) Using machine learning techniques for discovering latent topics in twitter colombian news. In: Narváez FR, Vallejo DF, Morillo PA, Proaño JR (eds) Smart Technologies, Systems and Applications, Springer International Publishing, Cham, pp 132–141

Vargas-Calderón V, Parra-A N, Camargo JE, Vinck-Posada H (2019) Event detection in colombian security twitter news using fine-grained latent topic analysis. 1911.08370

Williams NL, Ferdinand N, Bustard J (2019) From wom to awom–the evolution of unpaid influence: a perspective article. Tourism Review

Wong Ooi Mei A, Dean AM, White CJ (1999) Analysing service quality in the hospitality industry. Managing Service Quality: An International Journal 9(2):136–143

Xiang Z, Du Q, Ma Y, Fan W (2017) A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management 58:51 – 65, DOI https://doi.org/10.1016/j.tourman.2016.10.001

Xiang Z, Shin S, Li N (2019) Online tourism-related text: a perspective article. Tourism Review

Zeithaml VA, Bitner MJ, Gremler DD (2018) Services marketing: Integrating customer focus across the firm. McGraw-Hill Education,

Zhou S, Yan Q, Yan M, Shen C (2020) Tourists' emotional changes and ewom behavior on social media and integrated tourism websites. International Journal of Tourism Research 22(3):336–350

A Essentials of FastText

FastText is a library that creates text embeddings. This means that a string s is mapped to a vector in the vector space $\mathbb{R}^N$. The FastText method shares the embedding ideas from other models such as Word2Vec (Mikolov et al., 2013). In what follows, we will see the general ideas of how the map/embedding is built; however, the interested reader is referred to (Bojanowski et al., 2016; Joulin et al., 2016; Mikolov et al., 2013; Vargas-Calderón et al., 2019) for a more formal exhibition of the method with mathematical details.

Consider a dataset of texts, or documents $\mathcal{D}$. We can create the vocabulary set $\mathcal{V}$ as the set of words contained in the documents. We can order this set arbitrarily, but for the sake of simplicity, let us assume that we deal with a vocabulary that is alphabetically ordered. Let $V = |\mathcal{V}|$ be the size of the vocabulary. Consider a one-hot encoding map $\phi : \mathcal{V} \to \mathbb{R}^V$ be defined as a function that takes the i-th element of the vocabulary (in alphabetical order) and maps it to a vector ϕ_i, which has all of its components equal to 0 except the i-th component, which is equal to 1. The embedding is an $N \times V$ matrix W that maps a vector from the one-hot encoding vocabulary in $\mathbb{R}^V$ to the embedded vector space $\mathbb{R}^N$, where $N \ll V$. This means that the i-th word of the vocabulary will have an embedded vector representation $\boldsymbol{w}_i := W\phi_i$ (note that $\boldsymbol{w}_i$ is just the i-th column of W). The main feature is that words that are semantically similar, also have similar vector representations in the embedded space, i.e. $\boldsymbol{w}_i \cdot \boldsymbol{w}_j/(||\boldsymbol{w}_i|| \, ||\boldsymbol{w}_j||) \approx 1$ for similar words $w_i, w_j \in \mathcal{V}$.

The question that immediately arises is: how can one measure semantic similarity? Mikolov et al. (2013) define semantic similarity with a prediction problem that has its origin in the distributional hypothesis of linguistics (Harris, 1954), which states that semantically similar words are used in similar contexts. For instance, the words "kindness" and "courtesy" are expected to have similar vector representations because they can be found in positive comments about hotel staff with similar contexts. The context is formally defined as the set of words that surround the word of interest, and the amount of words that are taken into the context is normally referred as the context size. The definition of context allows us to state the prediction problem that defines the semantic similarity: given a context around a word of interest w_i, can we predict that the word of interest is w_i? or, given a word of interest w_i, can we predict its context? These two questions are answered by the continuous bag of words (CBOW) and the skip-gram configurations of Word2Vec-like architectures, respectively.

As an example, let us consider the CBOW configuration. Consider a part of a sentence consisting of a word of interest w (we drop the sub-index) and a context of size 4: $w_1 \, w_2 \, w \, w_3 \, w_4$. In the CBOW configuration, we use the context words to predict the word of interest. This is done by averaging the vector representation of the context words, i.e. $\boldsymbol{w}_c = \frac{1}{4}\sum_{i=1}^{4} \boldsymbol{w}_i$. The prediction of the word of interest[2] is done by computing $W^T \boldsymbol{w}_c$, which should equal to the one-hot encoding of the word w. The matrix elements of W can be learnt through any minimisation algorithm of a loss function such as categorical cross-entropy, built by sampling pairs (word of interest, context words) and predicting words of interest given their context words.

FastText (Bojanowski et al., 2016) leverages this idea to learn sub-word information embeddings. Instead of dealing with a vocabulary of words, FastText considers a vocabulary of n-char chains. To understand this, consider a sentence which contains the word "kindness". We use two special characters $\langle$ and $\rangle$ to mark where a word starts or ends, so that "kindness" is transformed to "$\langle$kindness$\rangle$". If we consider 5-char chains, we would get the following decomposition of "kindness": {"$\langle$kind", "kindn", "indne", "ndnes", "dness", "ness$\rangle$"}. We learn a vector representation for each 5-char chain found in our vocabulary in the same fashion of context words, and now, the representation of a word is the average of the representation of the chains that form its decomposition. This can be extended to sentence representation by also averaging its word representations.

---

[2] Here the prediction is made with the same weight matrix W. However, in practice, the prediction matrix has different weights, meaning that there are two different vector representations for each word.

B C_V topic coherence

The C_V topic coherence (Röder et al., 2015) is a metric that correlates well with human topic ranking, which gives a gold standard of interpretability. The C_V coherence is calculated as follows. For each topic, consider the set $W = \{w_1, \ldots, w_N\}$ of the N most frequent words within the documents assigned to that topic. Compute $p(w_i)$ as a frequency that tells the probability of finding word w_i in te documents of that topic. Also, compute $p(w_i, w_j)$ of finding w_i and w_j within a document, with the constrain that w_j must be at most s tokens away from w_i, where s is some fixed window size.

Now, we consider a segmentation of W, in the sense of Douven and Meijs (2007). Such a segmentation is a set of pairs of subsets of W. In particular, the C_V coherence uses a segmentation of the form

$$S = \{(W'_\beta, W) \,|\, W_\beta \in \mathscr{P}(W) - \{\varnothing\}\}, \tag{1}$$

where $\mathscr{P}(W)$ is the power set of W. We refer to each pair in S by $S_\beta = (W'_\beta, W)$.

We can represent each set of vectors $\bar{W} \in \mathscr{P}(W) - \{\varnothing\}\}$ with a context vector $\boldsymbol{v}(\bar{W})$ of size $|W|$, whose j-th component is

$$v_j(\bar{W}) = \sum_{w_i \in \bar{W}} \mathrm{NPMI}(w_i, w_j)^\gamma \tag{2}$$

, where where NPMI stands for normalised point-wise mutual information and γ assigns greater values to larger NPMI's. The NPMI is defined via

$$\mathrm{NPMI}(w_i, w_j)^\gamma = \left(-\frac{\log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}}{\log(p(w_i, w_j) + \epsilon)} \right)^\gamma, \tag{3}$$

where ϵ is a parameter added for numerical stability. Notice that the numerator in eq. (3) is just (ignoring ϵ) $p(w_i|w_j)/p(w_i)$, which will be greater than 0 if the conditional probability of w_i given w_j is greater than the probability of the word w_i. Therefore context vectors represent the level of co-ocurrence of a set of words $\bar{W}$, with respect to all words W in the N most frequent words within the documents of a topic.

Now, for each pair S_β, we compute a confirmation measure $\phi(S_\beta)$ (Syed and Spruit, 2017) as

$$\phi(S_\beta) = \frac{\boldsymbol{v}(W'_\beta) \cdot \boldsymbol{v}(W)}{||\boldsymbol{v}(W'_\beta)|| \, ||\boldsymbol{v}(W)||}. \tag{4}$$

The confirmation measure tells how strongly W supports W', i.e. how much semantically words from W are related to W' irrespective of how much two words (or sets of words) appear together in the corpus (see the work by Röder et al. (2015) for more detail on this). The average over all pairs S_β are taken as the coherence for the specific topic under study. Further averaging over all topics, gives the C_V coherence.