INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Sensing the impact of COVID-19 restrictions from online reviews: the cases of London and Paris unveiled through Text Mining

Bruno Rafael Martins da Silva

Master Degree in Telecommunications and Computer Engineering

Supervisor:
PhD Sérgio Miguel Carneiro Moro, Associate Professor with Habilitation,
ISCTE-IUL

Co-Supervisor:
PhD Catarina Maria Valente Antunes Marques, Assistant Professor,
ISCTE-IUL

December 2021

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Sensing the impact of COVID-19 restrictions from online reviews: the cases of London and Paris unveiled through Text Mining

Bruno Rafael Martins da Silva

Master Degree in Telecommunications and Computer Engineering

Supervisor:
PhD Sérgio Miguel Carneiro Moro, Associate Professor with Habilitation,
ISCTE-IUL

Co-Supervisor:
PhD Catarina Maria Valente Antunes Marques, Assistant Professor,
ISCTE-IUL

December 2021

Dedico esta dissertação à minha família, aos meus amigos e a todos os que contribuíram para a sua realização.

Acknowledgements

I thank my family for the strength given since the beginning of this long journey, to my friends for all their support and consideration throughout this period and especially to my sister who supported me a lot during this final stage. Without them I would not have the confidence to proceed with this project of personal achievement and knowledge building.

I thank my supervisors for all the tips, ideas and suggestions given during this last year and essentially for the vote of confidence they placed in me to deliver this work. Without a doubt, I managed to learn a lot from all the knowledge they shared.

To the colleagues from ISCTE who crossed paths with me in all these years I thank all the help and support, and to my job colleagues for all the comprehension in the times I needed to dedicate myself to the study and completion of this dissertation thesis.

Thank you all.

Resumo

Este estudo tem como objetivo compreender como a pandemia COVID-19 afetou o setor hoteleiro e identificar as exigências atuais dos hóspedes. As avaliações destes foram analisadas com base na análise de sentimento tendo sido usado para o efeito o website *TripAdvisor*, um dos sites mais populares na área do turismo, para a coleta de avaliações de hotéis em Londres e Paris. A coleta consistiu num *dataset* de 8k avaliações correspondentes a 226 hotéis. Os dados, em formato de texto, foram extraídos das revisões feitas em dois períodos homólogos, antes e durante a pandemia de COVID-19, para comparar o sentimento e os aspetos específicos destacados pelos hóspedes, entre cada período. A análise efetuada poderá igualmente servir como base para outros estudos que visem uma melhor compreensão do comportamento dos hóspedes na fase da pandemia de COVID-19.

Palavras-chave: *Text mining*, Análise de Sentimento, Turismo, Análises online de hotéis de viajantes, pandemia COVID-19.

Abstract

This study aims to understand how the COVID-19 pandemic affected the hotel sector and to identify the current traveler demands. The traveler's reviews were analyzed based on a sentiment and guest satisfaction analysis, demonstrating a data mining approach within tourism and hospitality research. Given its popularity, TripAdvisor was the chosen platform for collection of hotel reviews in London and Paris. The data collection consisted in a dataset of 8k reviews from 226 hotels in total. Text data were extracted from reviews made in two time periods, before and during the COVID-19 pandemic. The sentiment and specific aspects highlighted by travelers were compared between each period. This analysis may also serve as a basis for other studies aimed at a better understanding of the behavior of hotel guests during the COVID-19 pandemic.

Keywords: Text mining, Sentiment Analysis, Tourism, Hotel traveller's online reviews, COVID-19 pandemic.

Index

Acknowledgements iii

Resumo v

Abstract vii

CHAPTER 1. Introduction 11

1.1. Research question 11

CHAPTER 2. Literature Review 13

2.1. Mining from online reviews 13

2.1.1. Web Scraping 14

2.1.2. Sentiment Analysis 14

2.2. Data Mining applied to global tourism and the impact of COVID-19 15

CHAPTER 3. Methodology 19

CHAPTER 4. Results and discussion 21

4.1. Analysis on TripAdvisor reviews 21

4.1.1. Guests' nationality perspective 22

4.1.2. Hotels' rating 23

4.1.3. Traveling type analysis 24

4.1.4. Most used words 25

4.1.5. Sentiment Analysis 27

CHAPTER 5. Conclusions 33

References 35

Attachments 37

A - Scientific Article 37

CHAPTER 1

Introduction

Over the past years, we have witnessed a notorious increase on the number of text reviews made in several online accommodation platforms following the exponential growth in tourism (Chan et al., 2021). However, during the year of 2020 as COVID-19 pandemic continued to spread, this has changed completely, namely with the drastic decrease in demand of travel research (Uğur & Akbıyık, 2020).

COVID-19 pandemic brought a radical change in people's lives with many economic sectors severely damaged (Chan et al., 2021). Tourism was one of the most affected business sectors contrasting with the steady growth seen over the years (Chan et al., 2021). For this reason, it is now extremely important to help tourism, and more specifically hotels, to understand how COVID-19 affected the sentiment of travelers during their stays.

The present research aims to understand how the current COVID-19 pandemic restrictions affected tourism and the feelings of travelers using text mining, namely sentiment analysis, on hotel traveler's online reviews. The TripAdvisor reviews were collected from two capital cities, London and Paris, for being very similar in terms of tourism characteristics which fulfilled the two main objectives of the study: one is the identification of a gap on subareas within tourism on which data science has not yet been applied, and the second, to perform data mining on online reviews extracted from the TripAdvisor platform and submitted during COVID-19 pandemic. These goals will help to understand what travellers seek and what demands they have.

Furthermore, a scientific article related to this research was submitted and presented to the ICMarkTech 21 international conference, as well as published in the Proceedings by Springer and available in the SpringerLink Digital Library. The article can be consulted in the "Attachments" section of this document.

1.1. Research question

This research falls into the current COVID-19 pandemic focused on how this affected tourism and the feelings of travellers.

Tourism was one of the most affected business sectors contrasting with the steady growth seen over the years (UNWTO, 2021). Text mining is one of the most useful methods of big data analytics due to its ability to filter and retrieve specific topics and words (Hassani et al., 2020). This can help hotels to better understand their clients and to implement options and measures to fulfil their demands. In the end, hotels, and tourism in general will be able to improve their profits on such a critical situation this sector is living nowadays (UNWTO, 2021).

The developed research focused on answers essentially for one question: “How is the effect of the pandemic being perceived by visitors to major European capitals severely affected by COVID-19?”.

During the study and following the answer for this question some hypotheses were tested, described, and the discussion included later in this document.

CHAPTER 2

Literature Review

2.1. Mining from online reviews

On a recent study on an analysis of TripAdvisor reviews published by (Sangkaew & Zhu, 2020), it was stated that nowadays tourists have taken the initiative to share their travel experiences on multiple channels, such as Facebook, Yelp, Instagram and TripAdvisor. Online reviews not only allow tourists to record their experiences and emotions, both positive and negative but also act as a significant and easily way to provide valuable information for other tourists during their destination planning and decision-making process. Moreover, online reviews offer valuable insights and feedback for local tourism stakeholders to understand how tourists perceived tourist attractions, experiences, and destinations to a broader extent (Sangkaew & Zhu, 2020). Most online reviews are generated because tourists want to share their genuine personal experiences with a wider audience, leaving the content immediately accessible for data collection as well as unbiased and uncontaminated for credibility and trustworthiness (Sangkaew & Zhu, 2020).

Some other studies have also analysed online reviews to explore tourists' experiences in various contexts, such as glamping, sports events and medical tourism. Many scholars have also investigated various topics in the tourism and hospitality context by using online content and reviews as their source of data. Many studies have examined complaints about hotels by means of online reviews which not only allowed to record the actual experiences of tourists and deliver messages to potential tourists, but also created the potential to be used as valuable feedback from tourists to learn about their satisfaction level and upgrade tourism product and service (Sangkaew & Zhu, 2020).

Data Mining refers to the process of advance analysis of extensive data sets. This analysis can be advanced enough to require machine learning technologies to uncover specific trends or insights from the dataset (Perez, 2020). For example, data mining might be used to analyse millions of transactions from a retailer such as Amazon to identify specific areas of growth and decline. In some cases, web scraping might be used to extract and build the data sets that will be used for further analysis via Data Mining (Perez, 2020).

Data mining helps researchers to investigate unsuspected relationships in the data and to provide useful insights to the data owners, so that when data mining techniques are used, they can give some new and essential insights that have never been discovered before by researchers and practitioners (Chen et al., 2020). Data mining has been applied for different tasks, especially for analysing text, including online review comments. In recent years, major statistical software and data mining programs, such as text mining functions, have been developed to speed up the analysis process of unstructured data, including email, text comments, web documents, pictures, and images. Analysing data from online customer reviews by using data mining methods can provide meaningful insights into service performance (Chen et al., 2020). Data mining approaches can deal with big data, while the traditional statistics techniques cannot easily handle large databases efficiently. In the work of (Guo et al., 2017), the authors analysed 266,544 online reviews extracted from 25,670 hotels to recognize crucial dimensions of customer services based on data mining methods. Lim & Lee (2019) used data mining methods to analyse passengers' online comments regarding airline services. Thus, using data mining to analyse customer satisfaction from online reviews is effective (Chen et al., 2020).

2.1.1. Web Scraping

Web scraping refers to the process of extracting data from web sources and structuring it into a more convenient format. It does not involve any data processing or analysis. Whilst web scraping can be done manually, in most cases web scraping software tools are preferred due to their speed and convenience. In fact, web scraping could be used to create the datasets to be used in Data Mining (Perez, 2020).

2.1.2. Sentiment Analysis

Sentiment analysis is a natural language processing tool that is useful for monitoring Web 2.0 applications, as it can reveal public opinion about numerous issues without requiring satisfaction enquiries (Valdivia et al., 2017). According to a published article about sentiment analysis from data extraction on tweets (Preethi et al., 2015), "it is the process of computationally identifying and categorizing opinions expressed in a piece of text to determine whether the writer's attitude toward a particular topic, product, and so on is generally positive, negative, or neutral". The interest in sentiment analysis has increased significantly over the last few years due to the large amount of text stored in Web 2.0 applications and the importance of online customer opinions (Valdivia et al., 2017). As a result, more than 1 million research papers contain the term "sentiment analysis" and various start-ups have been created to analyse sentiments in social media companies (Valdivia et al., 2017).

2.2. Data Mining applied to global tourism and the impact of COVID-19

The spread of the COVID-19 global pandemic has generated an exponentially mounting and extraordinary volume of data that can be harnessed to improve our understanding of big data management research (Sheng et al., 2020). It is also applied in the necessity among scholars, practitioners, and policymakers for a better and deeper understanding of a range of analytical tools that could be utilized to better anticipate and respond to such unforeseen 'black swan' events and risks. Indeed, advancements and proliferation of different technologies have culminated in unprecedented production of mobile, digital devices and a vast amount of structured and unstructured data to be mined by firms and governments for sound and timely decision-making (Sheng et al., 2020).

Although there has been amassing of both unstructured and semi-structured data across the globe on such exogenous shocks, much of the current growing data remains untapped, to the detriment of wider society and policy (Sheng et al., 2020). As recently observed by The Economist, 'the world's most valuable resource is no longer oil, but data (Sheng et al., 2020). The issue of the world's under-utilized asset is exacerbated by a growing number of methodological approaches, but we lack a deeper understanding of the different techniques and how some can be utilized in concert to improve researchers' approaches and tackle new global issues such as COVID-19 (Sheng et al., 2020). Predictive analytics concerns what will happen in the future and is generally considered as the use of "statistical techniques to analyse current and historical facts to make predictions about future events and/or behaviour" (Sheng et al., 2020). One of the categories classified within predictive analytics are methods for analysing unstructured data like text mining that we are going to cover in the current research.

Chan et al. (2021) that revealed the differences in guest experiences at luxury hotels before and during the pandemic as reflected in the online reviews posted in those two periods. A text analytic approach was used to transform unstructured data and identify factors in online reviews to compare the determinants of hotel guest experiences before and during the pandemic. The results offered practical suggestions to hotel managers to develop effective operational and marketing strategies to improve guest experiences at luxury hotels during the challenging situation. This research focused on textual reviews in hospitality that identified hotel attributes and sentiments expressed in the reviews and examined the relationships between specific attributes in the textual reviews and the overall review ratings. Hong et al. (2020) found that guests placed more emphasis on natural and safe experience associated with the bed & breakfast (B&B) after COVID-19 and provided practical suggestions for the industry to survive the disaster, such as avoid using central air-conditioning and adopting semi-self-service technologies (Chan et al., 2021).

Online reviews for 20 luxury hotels in Shanghai that were posted from August 2019 to July 2020 were collected using 'Google Sheets' in August 2020. These 20 hotels were randomly selected based on the list of luxury hotels in Shanghai that was available on Ctrip (Chan et al., 2021). Mainland China announced the lockdown measures to control the spread of COVID-19 around the country in late January 2020. Online reviews posted from August 2019 to January 2020 were categorized into the pre-COVID-19 group (740 reviews), while those posted from February 2020 to July 2019 were in the amid-COVID-19 group (1238 reviews) (Chan et al., 2021). There was a significant difference in the number of reviews between these two periods. The distribution showed that a very small number of reviews has been posted in September and October 2019 before the outbreak of the pandemic. However, a lot of reviews have been posted in June and July 2020 during the pandemic. This suggested that consumers have a higher motivation to share their experiences with and help the others amid-crisis. Table *1* shows a summary of the method and findings of study published by Chan et al. (2021).

Table 1. Research on differences in guest experiences at luxury hotels before and during the pandemic

Reference	Goal	Method	Findings
(Chan et al., 2021)	Compare the determinants of guest experience at luxury hotels in Mainland China before and during the pandemic—COVID-19.	Online reviews for 20 luxury hotels in Shanghai that were posted from August 2019 to July 2020 were collected using Google Sheets in August 2020. The online reviews were analyzed using the Chinese Text Analyzer software, which helps to perform segmentation and frequency count of the words.	Hotel guests mentioned much more about services during the pandemic (21.8%) compared to the time before (17.3%). The proportion of reviews mentioning about health measures increases from 0.8% before the outbreak of the pandemic to 3.2% during the pandemic.

2.3. Conceptual framing and research hypotheses

Based on other studies that identified a list of dimensions drawn from quantitative features known to influence customer satisfaction under the context of TripAdvisor, the list below (Table *2*) was considered to support a data mining analysis and approach within tourism and hospitality research.

Table 2.List of considered dimensions for our study

Dimension	Reference
Services	(Chan et al., 2021)
Amenities	(Moro et al., 2019)
Health measures	(Chan et al., 2021)
Hotel facilities	(Moro et al., 2019)
Location	(Chan et al., 2021)
Value/price	(Chan et al., 2021)
Cleanliness	(Moro et al., 2019)
Type of travel	(Moro et al., 2019)
Hotel prestige	(Moro et al., 2019)
Seasonality	(Moro et al., 2019)

Restuputri et al. (2021) published a research comparing the staff service quality, operational, technical logistics service providers, with customer satisfaction and loyalty during the COVID-19 pandemic. In this study they refer that an employee must be reliable, punctual and careful at work. An employee must also have effective communication skills, be courteous and ready to serve. The quality of operations service from source to customers must be well-coordinated, on time and with appropriate transportation capacity so no damage occurs to the customer's property. In the end they concluded that a good relationship between staff, operational and technical services is crucial for a good perception of customer satisfaction and loyalty to the services provided by hotels.

Big crisis affects the macro-environment that brings big changes in customer behaviors and hotel performances (Chan et al., 2021). The outbreak of COVID-19 has brought changes in guest experiences at hotels as customers usually describe their experiences and feedbacks after their stays in the form of user-generated content, such as online reviews which are then used by practitioners to understand the nature and structure of guest experiences (Chan et al., 2021). Since hotels are customer-centric that should keep up with customer preferences and requirements, it is essential to understand the impact of COVID-19 and track the changes of guest experiences brought by the same (Chan et al., 2021).

Based on the above studies, the following individual hypothesis were tested in this research:

H1: Hotel guests appreciate safety concerns by hotels and express about them in online reviews.

Another study published by Uğur & Akbıyık (2020) mentions the result of extracted phrases that did outstand on the number of repetitions such as travel insurance as the second most frequently repeated phrase. This phrase refers to protecting travellers from trip cancellation to flight delays and ensuring assistance for medical emergency or luggage loss (Uğur & Akbıyık, 2020). Also, cases containing credit card, full refund, and travel agency were examined and it was observed that the comments were about the refund of all or some part of the payments and change or cancellation of travel plans. There were also included phrases regarding pandemic protection measures such as wear mask, face mask and hand sanitizer (Uğur & Akbıyık, 2020). This suggested two additional hypotheses considered to be tested in our study:

H2: The need of travel insurance has raised with the pandemic.

H3: The existence of a full refund option in case of travel cancelling.

CHAPTER 3

Methodology

The method used for this research is based in a problem resolution paradigm that contains several activities (Alexandre, 2017):

Problem identification and motivation was the first step which defines that the artefact must be viable, that is, it must provide a solution to the research problem. In the presented case, it will be produced in the form of a method to study the impact of COVID-19 restrictions on the TripAdvisor reviews from 2 foreign locations very similar in terms of tourism characteristics. TripAdvisor is an American travel website company providing reviews from travellers about their experiences in hotels, restaurants, and monuments (Valdivia et al., 2017). It has become the largest travel community, reaching 390 million unique visitors each month and listing 465 million reviews and opinions on more than 7 million accommodations, restaurants, and attractions in 49 markets worldwide (Valdivia et al., 2017).

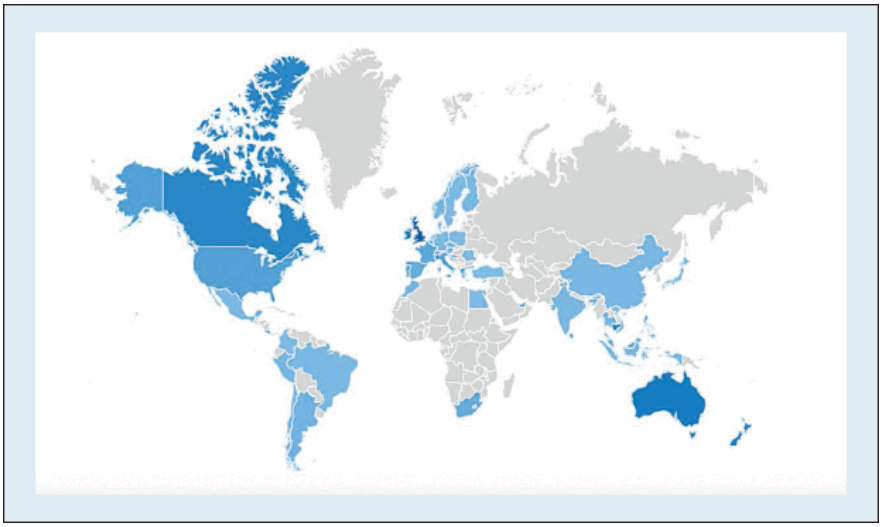

Figure 1. Map of the popularity of TripAdvisor Google searches (2012 to 2017).

TripAdvisor emerged in 2004 as a Web 2.0 application for the tourism domain. This user-generated content website offers a plethora of reviews detailing travellers' experiences with hotels, restaurants, and tourist spots. TripAdvisor has since been ranked as the most popular site for trip planning, with millions of tourists visiting the site when arranging their holidays. Since it has so much data, TripAdvisor has become extremely popular with both tourists and managers (Valdivia et al., 2017). Tourists can read the accumulated opinions of millions of everyday tourists. They can also check the popularity index, which is computed using an algorithm that accounts for user reviews and other published sources such as guidebooks and newspaper articles. This index runs from number 1 to the overall total number of restaurants, hotels, or other attractions within the city. Travelers can find the most interesting visitor attraction or the most popular restaurant. Linked to this is the bubble rating (user rating), a 1–5 scale where one bubble represents a terrible experience and five bubbles an excellent experience. All reviewers are asked to use this scale to summarize their feedback. Alongside with this rating, users include their opinions, which can cover the performance of a restaurant, hotel, or tourist spot (Valdivia et al., 2017).

Chosen the platform for the data extraction it was necessary to define the solution objectives, which must be achievable, in addition of the need to understand if this analysis could identify a research gap in the utilization of data mining in tourism.

The third step, Development, involved the current creation of the artefact defined in the first step, that is, the development of a method that allows the analysis and comparison of both hotel reviews. For the sake of the extraction a pre-existent coding script was chosen to perform the scraping, that contained several features to quickly setup and extract data from a website. In this phase it was necessary to understand if the objective and its influencing factors could be translated into data features and instances (Moro et al., 2019). As soon as this was achieved, the data also needed to be prepared (data cleaning) for model training.

Next to development was the Demonstration, which was used to verify the effectiveness of the solution to the problem identified. To achieve this goal, graph plots and word clouds or any other most appropriate visualization methods were used in addition to and a pre-build script to plot the most common words from the hotel reviews.

The qualitative measurement of how well the created artefact supports the solution to the problem corresponded to the next step, the Evaluation, which involved the use of the model previously trained to make predictions versus observations (Moro et al., 2019).

And lastly, the communication of the results, whose purpose helped highlight and identify the importance of the problem through the dissemination of the developed artefact, that is, its usefulness and its relevance to other researchers (Alexandre, 2017).

CHAPTER 4

Results and discussion

4.1. Analysis on TripAdvisor reviews

The following output results are demonstrated mainly with word cloud plots, charts in addition to tables and/or figures every time it was required. This helped to clarify of what were the major differences within the customer reviews between the pre and during COVID-19 restrictions on equivalent time periods. In addition, a comparison between two major capital cities was provided.

Figure *2* represents the number of reviews extracted per year and city. A significant reduction can be immediately seen from 2019 to 2020 in both cities. In case of London there is, in average, a reduction of 86% on the reviews and Paris with a 95% of decrease. Nevertheless, on 2020, London had a slight increase between June and October mainly due to ease of lockdown restrictions in that period (Institute for Government, 2019), followed then by a second decrease as the number of deaths from Covid-19 disease were rapidly rising, as per Figure *3*.

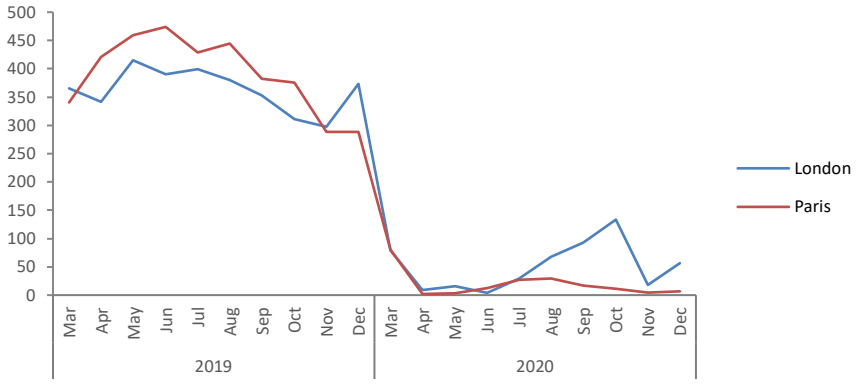

Figure 2. Distribution of reviews during 2019/2020 periods

Figure *3* and Figure 4 provide a more detailed insight about the evolution of reviews of TripAdvisor (left Y axis) and the number of Covid-19 cases and deaths (right Y axis), respectively in each figure, reported in each city per month. Figure *3* shows the data from 2019, to have a fair comparison in the homologous period. In 2020, an increase in the number of reviews, in London, can be seen during the peak of Summer. An increase is also noticeable in Paris but with much less evidence. This difference is related to the severity of the restrictions implemented on each city which were higher in Paris.

In fact, the charts below (Figure *3* and Figure 4) allow to determine the periods where each government applied and eased the pandemic lockdown restrictions. In the case of London, the ease of

restrictions started on 23rd of June 2020 and, on 31st of October 2020 England announces second national lockdown (Institute for Government, 2019). During this period the number of reviews done on TripAdvisor rose compared with the first wave of cases.

On a same analogy, in Paris, the government eased the restrictions on 14th of June 2020. On 20th of August France announced that the pandemic was again on the rise (Aurore & Blue, 2020). Due to this, there was also a slight increase in TripAdvisor reviews but with much less expression, compared to London.

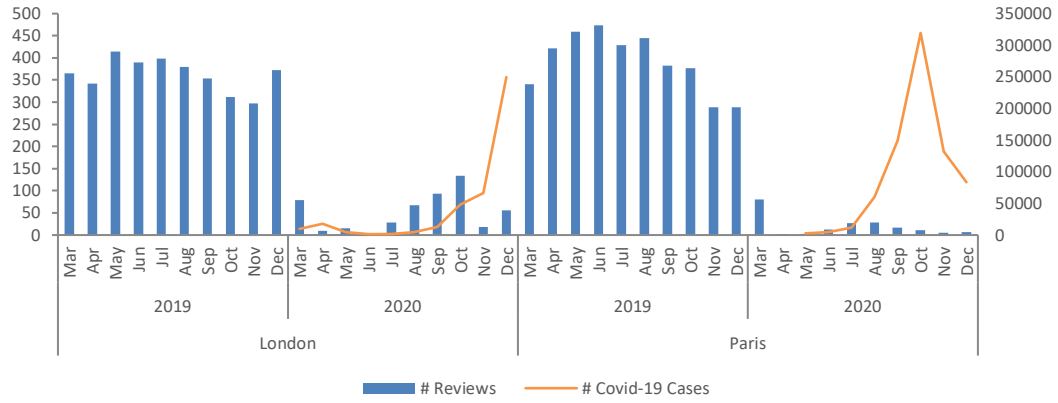

Figure 3. Distribution of reviews (2019 vs 2020) and Covid-19 cases (2020)

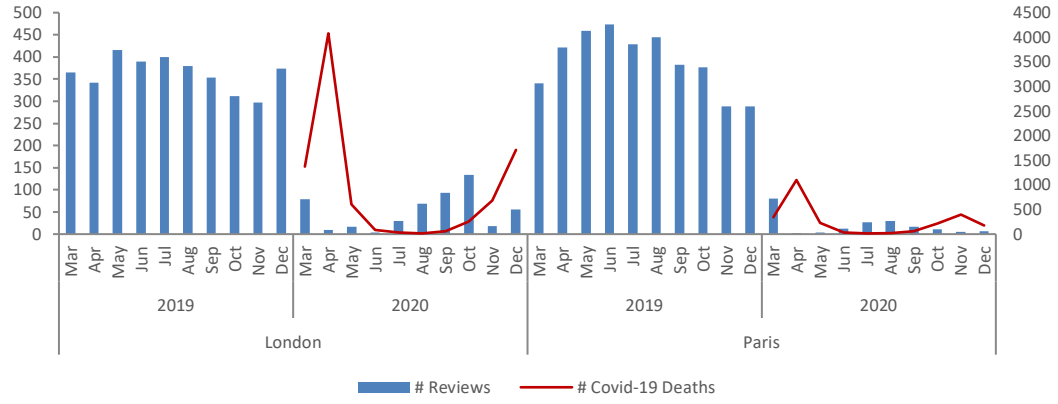

Figure 4. Distribution of reviews (2019 vs 2020) and Covid-19 deaths (2020)

4.1.1. Guests' nationality perspective

Figure *5* shows the difference in terms of the nationality of the TripAdvisor reviewers. The total number of reviewers categorized by nationality and the percentage based on the total in each year are compared. The scenario of London, in 2020, is an example of expected decrease in the total number of reviewers (87%) and when comparing in terms of nationality a significant increase is noticeable on the national tourists of about 40%.

Paris, on the other hand, demonstrates that the city is more visited by the international tourists than those living within borders, in 2020. This behaviour could be explained by the fact of the collected data being filtered in the English language only. Also, this last point can explain the significant decrease in the written reviews on TripAdvisor (95%) as it is very likely that, for Paris, there are even more reviews written, by national guests, that is reviews written in French.

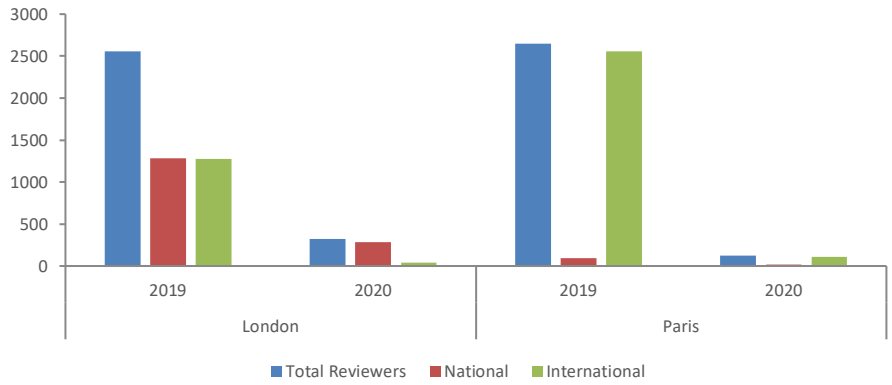

Figure 5. Comparison between National and International reviewers per city and year

4.1.2. Hotels' rating

During the data scraping it was also extracted the average rating value of the reviews so that some conclusions could be done. TripAdvisor rating values consist in a bubble rating system from one bubble to five bubbles, with one bubble meaning "terrible" and five bubbles meaning "excellent" (*Tripadvisor Help Center*, 2021.). Figure 6 presents the evolution of hotels rating during the 2019-2020 period per city. The bars show only the ratings above or equal to three bubbles since the data collected with ratings lower to this value were very small, and the solid line represents the average review rating from the stay of the guest. In the case of London, on a pre-pandemic scenario, the average rating given was four bubbles. During the pandemic period, in 2020, that average got a visible increase as it can be noticed.

On the other hand, Paris had a significant decrease between April and May of 2020, being this caused by a very low number of reviews in that time. From that point onwards it is possible to see an increase tendency on the average rating value. This positive tendency, in 2020, could be explained due to high demand of quality on the stay of the guest during the pandemic period. The feeling of safety and cleanliness might be also a reason. In London, this is also noticeable by the higher percentage of 5-star hotel preference.

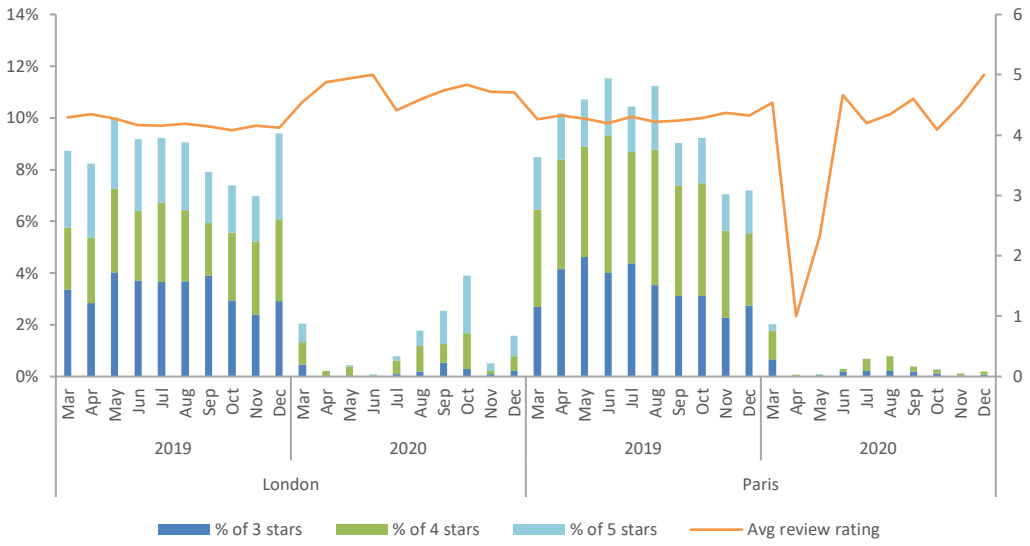

Figure 6. Distribution of hotel stars per city and year over total number of reviews

4.1.3. Traveling type analysis

An analysis on the type of travellers was also performed as per Figure 7.

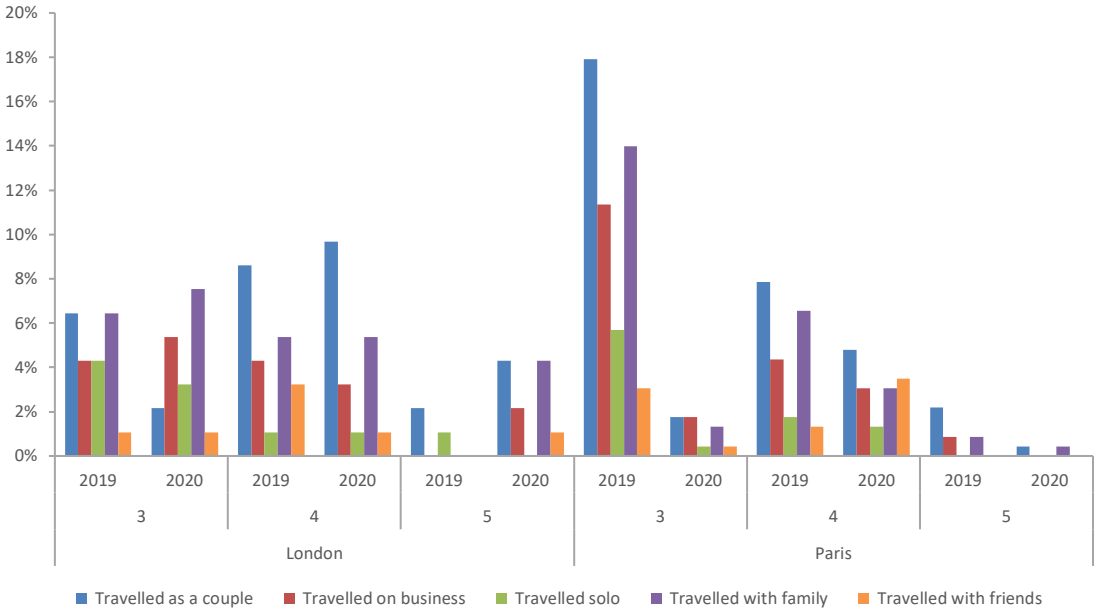

Figure 7. Distribution of type of travel per city, hotel stars and year over total number of reviews

There is a relevant choice for the 3-star hotel category in Paris during 2019 which significantly decreased in 2020 replaced then by the 4-star hotel. The demand on services quality is again demonstrated. London on the other hand has the choice being taken by the 5-star category, on 2020, with the 3 and 4-star categories not having a significant variation.

During the pandemic period, the most frequent type of travel in London were registered by the couples, followed by family travels. In Paris, couples register most part of the visits as well, in 2019.

This could be explained by the fact that people are more confident on travelling with their own family and couple members unlike travelling with friends only which was the least reported type of travel. Table 3 shows in more detail the distribution values for a clearer understanding of the analysis.

Table 3. Distribution of type of travel per city, hotel stars and year over total number of reviews

City	Hotel stars	Year	Travelled as a couple	Travelled on business	Travelled solo	Travelled with family	Travelled with friends
London	3	2019	6.45%	4.30%	4.30%	6.45%	1.08%
		2020	2.15%	5.38%	3.23%	7.53%	1.08%
	4	2019	8.60%	4.30%	1.08%	5.38%	3.23%
		2020	9.68%	3.23%	1.08%	5.38%	1.08%
	5	2019	2.15%	0.00%	1.08%	0.00%	0.00%
		2020	4.30%	2.15%	0.00%	4.30%	1.08%
Paris	3	2019	17.90%	11.35%	5.68%	13.97%	3.06%
		2020	1.75%	1.75%	0.44%	1.31%	0.44%
	4	2019	7.86%	4.37%	1.75%	6.55%	1.31%
		2020	4.80%	3.06%	1.31%	3.06%	3.49%
	5	2019	2.18%	0.87%	0.00%	0.87%	0.00%
		2020	0.44%	0.00%	0.00%	0.44%	0.00%

4.1.4. Most used words

The implemented algorithm had also the possibility to determine the most common words retrieved from the guests' reviews. Figure 8 contains the top 25 most frequent words in each of the cities.
The topmost common words used are "hotel", "room", "stay" and "staff", in both cities of London and Paris and in both years. Other words such as "good", "great" and "clean" are also commonly referred meaning that in general the stay of the guest had a positive feedback.

Below, on Table 4, another type of overview of the topmost common words extracted is shown. The biggest words, in size, are the ones mostly referred and the smallest are the not so common.

In terms of words frequency, no major difference was noticeable between the two periods, before and during the pandemic, suggesting that guests did not change much their behaviour and perspective related to the hotels in general.

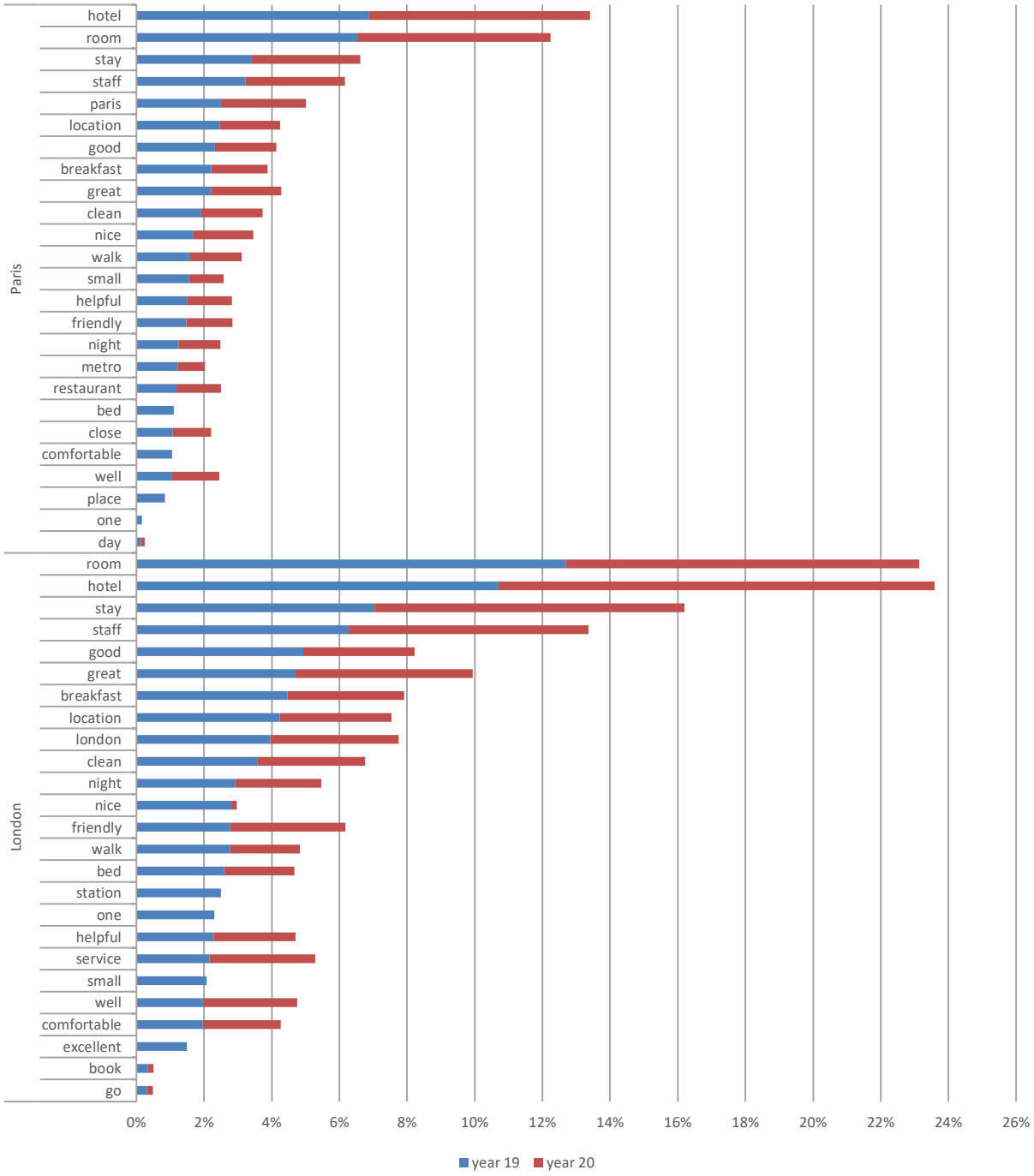

Figure 8. Overview of the top 25 most common words per city and year over total number of reviews

Table 4. Overview of the topmost common words per city and year showed in a 'Wordcloud' perspective

	London	Paris
2019		
2020		

4.1.5. Sentiment Analysis

The sentiment analysis part was also included in the mining process of TripAdvisor's extracted data. Figure 9, Figure 10 and Figure 11 present three different perspectives on the analysis of guest satisfaction based on the dimensions chosen for the study that were already discussed in previous chapters.

To evaluate the guest satisfaction, in each review, the sentiment words were drawn, and their scores were calculated. A score higher than zero was considered a positive review, a score equal to zero a neutral score and finally a score less than zero a negative one. For example, reviews with the words "helpful", "greeting", "friendly" and "amazing" got the best scores and reviews with word combinations like "excellent space optimization" and "real pleasant surprise" also were classified with best positive scores. After the key words' identification for each review, a new classification was done using the list of ten dimensions shared in the study. An average was then calculated per each dimension based on positive and negative perspectives.

As per Figure 9 and Figure 10 almost every dimension was classified with a positive sentiment, in both years. It is possible to see that in Paris during 2020, seasonality negatively impacted the reviewers when scoring their stays. This may be justified by the pandemic situation lived during this time. Health

measures is another dimension affected by this reason getting 42% of negative reviews. In the case of London, curiously, was not possible to retrieve any feedback referring both seasonality and health measures dimensions.

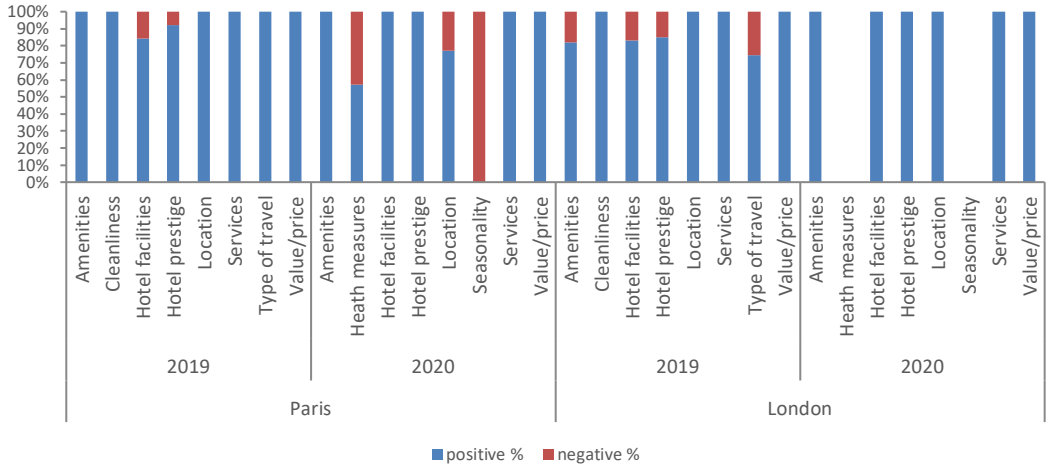

Figure 9. Sentiment perception distribution by dimension over total number of reviews

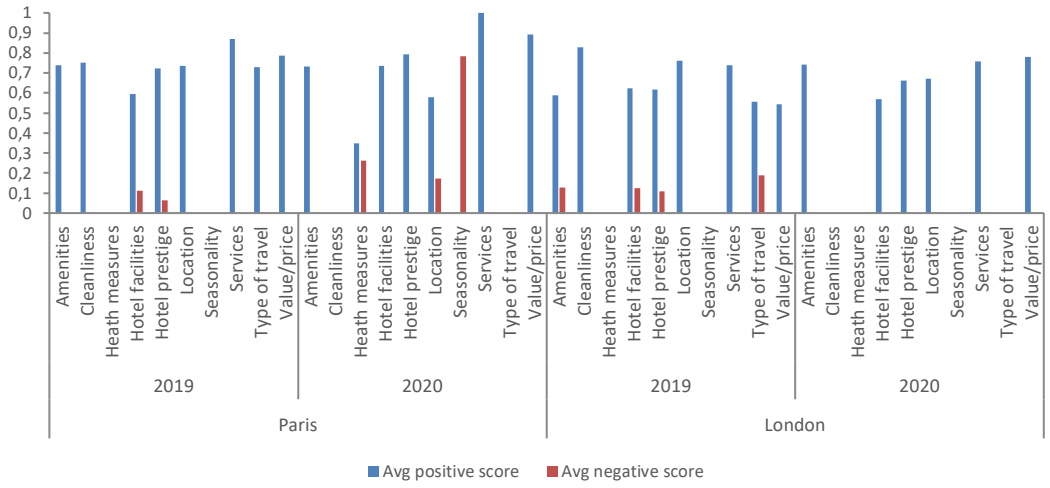

Figure 10. Sentiment perception average score by dimension, city, and year over total number of reviews

Figure 11 shows another perspective of the dimension analysis. As already referred seasonality clearly stands out as the most negative perceived dimension, in 2020. Health measures were also classified with low review scores. In 2019 the dimensions with poorer review sentiment were related to amenities, hotel facilities and prestige and the type of travel of the guests.

In the other hand, dimensions like the services and value/price were the ones with better sentiment classification, in both cases of Paris and London.

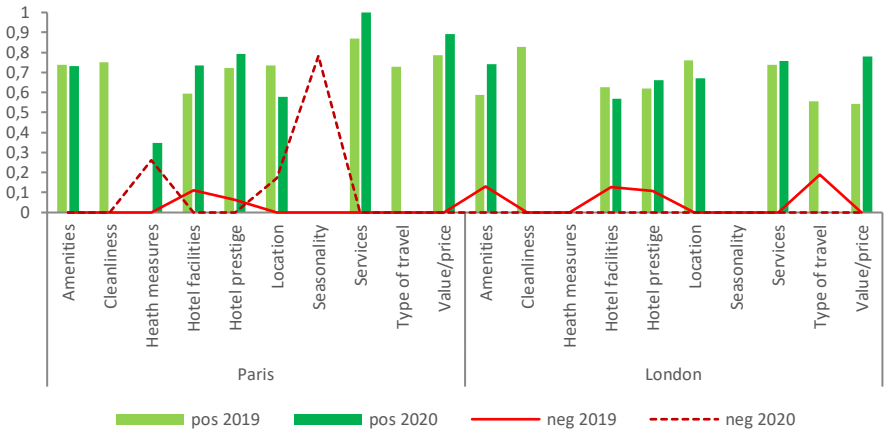

Figure 11. Sentiment perception by dimension and year over total number of reviews

Table 5, on next page, contains a summary analysis of each of the dimensions discussed as well as a satisfaction rank for a better understanding on how each dimension was perceived by guests in both pre and pandemic periods. A positive evolution is represented with a green highlight, a negative with red and the neutral, meaning that the dimension stayed in the same position from 2019 to 2020, represented with a blue highlight.

Table 5. Effects on customer satisfaction per dimension

Dimension	Effect on customer satisfaction		Satisfaction Rank	
	Paris	London	Paris	London
Services	In general, services have a positive effect on the guests; 2019 – the best rated dimension; 2020 – the worst rated.	The dimension has also a positive perceived value on customer; 2019 – best rated than in 2020 (with a neutral feedback).	2019: 1 **2020: 1**	2019: 3 **2020: 2**
Amenities	74% of positive feedback in both years. No negative effect perceived by guests.	2020 – the second-best rated dimension; 2019 – the second worst rated dimension. Nevertheless, the average positive rating is greater than the negative feedback.	2019: 4 **2020: 5**	2019: 6 **2020: 3**
Health measures	2019 – top negative rated dimension; 2020 – second negative rated dimension along with Cleanliness.	No feedback extracted.	2019: 9 **2020: 7**	2019: 9 **2020: 9**
Hotel facilities	Third negative dimension in 2019; Mostly positively rated in 2020.	Positive rating in general.	2019: 8 **2020: 4**	2019: 4 **2020: 6**
Location	Increase of 17% of negative rating value in 2020 compared to 2019.	100% positive feedback.	2019: 5 **2020: 6**	2019: 2 **2020: 4**
Value/price	Top positive rated dimension in 2020 and in overall.	Second best positive rated dimension.	2019: 2 **2020: 2**	2019: 8 **2020: 1**
Cleanliness	Second most negative rated dimension	Top positive rated dimension in 2019.	2019: 3 **2020: 8**	2019: 1 **2020: 7**
Type of travel	More decisive for travelling in 2019.	More decisive for travelling in 2019.	2019: 6 **2020: 9**	2019: 7 **2020: 8**
Hotel prestige	Mostly positively rated in 2020.	Positive rating improvement from 2019 to 2020.	2019: 7 **2020: 3**	2019: 5 **2020: 5**
Seasonality	Mostly negative perceived by guests in 2020.	No feedback retrieved from analyzed guests' reviews.	2019: 10 **2020: 10**	2019: 10 **2020: 10**

Based on the results from Table 5 it's possible to see that the feedback retrieved from the dimensions "Health measures" and "Cleanliness" take a relevant place into the most rated dimensions overall as both dimensions are most often the top or second highest rated in the positive and negative feedback

sides. This immediately helps to conclude that our first proposed hypothesis related to the hotel guests' appreciation of safety concerns by hotels and expression about them in online reviews was validated. Still, it's worth to mention that the case of London has greatly contributed for this achievement, when comparing to some lack of feedback extracted on Paris reviews.

Dimensions like services and value/price are the most well perceived by guests. Seasonality, on the other hand, was one of the worst decisive factors on the guests' opinions on to their stays.

Unfortunately, it was concluded that the data extracted wasn't detailed enough to test and corroborate the two remaining proposed hypotheses concerning the need of travel insurance and existence of a full refund option. This would require another level of data extraction such as analysing data from web sources related to flight companies, for instance.

CHAPTER 5

Conclusions

The main objective of this study was to understand what travellers are seeking nowadays during the current pandemic that we are all living. From the results and analysis done on the data mining process it was possible to understand some aspects of the behaviour, profile and most importantly how guests are now perceiving the effect of COVID-19 pandemic in their hotel stays in two major capital cities being this the major research question of this investigation.

In more detail, the results presented demonstrated that most of the reviews (55%) done on TripAdvisor on 2020 showed a positive perception, despite the COVID-19 pandemic when comparing to 2019. Moro et al. (2019) found in their published study that the most relevant dimensions concerning guests' satisfaction were the previous user's experience with the online platform, the individual preferences and hotel prestige. This last one was also considered and analysed in our study, but we can say that it was a "mid" rated dimension, that is, not being the most or worst perceived factor by the guests. In our case we found that the value/price and services dimensions were the most positively perceived by guests, meaning that hotels still, within a mid-pandemic crisis, need to offer a good quality balance in the services they provide. This was also corroborated by another important finding, related to the preference of guests on a high hotel star category during the crisis of the pandemic. On the other hand, the dimension related to the health measures implemented in the hotels was not so well perceived by guests in 2020, showing the constant concern around the preventive measures during the pandemic. Curiously, feedback about this dimension was only obtained in Paris. This could be justified by some technical limitations felt during the analysis and data extraction process meaning that some improvements may need to be implemented into the mining process.

Furthermore, since this study was based only in reviews written in English the number of data extracted was not equivalent in both cities and that could have limited some more exploration on the comparison between them. In addition, and as previously mentioned, it was not possible to corroborate and test the hypotheses regarding travel insurance and full refund option.

This dissertation thesis can also be the base of a continuity study in the future as many aspects most probably have changed and additional variables appeared since the start of the study. One of the those to be considered is the COVID-19 vaccination rate evolution on the globe that could have made a great impact in the hotel reviews. In addition, it would be worth to perform a deeper analysis into the TripAdvisor reviews data such as extracting the day guests book their travel. Despite not being a

mandatory field, it could lead for testing an additional hypothesis like if the fear of confinement requirement leads visitors to book their travels only on the same day they will actually travel.

References

Alexandre, I. M. (2017). Design Science Research. 1–19.

Aurore, J., & Blue, F. (2020). Coronavirus: the key dates of the epidemic in France. https://www.francebleu.fr/infos/societe/coronavirus-les-dates-cles-de-l-epidemie-en-france-1603646805

Chan, I. C. C., Ma, J., Ye, H., & Law, R. (2021). A Comparison of Hotel Guest Experience Before and During Pandemic: Evidence from Online Reviews. In Information and Communication Technologies in Tourism 2021 (pp. 549–556). Springer International Publishing. https://doi.org/10.1007/978-3-030-65785-7_52

Chen, W.-K., Riantama, D., & Chen, L.-S. (2020). Using a Text Mining Approach to Hear Voices of Customers from Social Media toward the Fast-Food Restaurant Industry. Sustainability, 13(1), 268. https://doi.org/10.3390/su13010268

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. Big Data and Cognitive Computing 2020, Vol. 4, Page 1, 4(1), 1. https://doi.org/10.3390/BDCC4010001

Hong, Y., Cai, G., Mo, Z., Gao, W., Xu, L., Jiang, Y., & Jiang, J. (2020). The Impact of COVID-19 on Tourist Satisfaction with B&B in Zhejiang, China: An Importance–Performance Analysis. International Journal of Environmental Research and Public Health, 17(10), 3747. https://doi.org/10.3390/ijerph17103747

Institute for Government. (2019). Lifting lockdown in 2021: the next phase of the coronavirus strategy | The Institute for Government. https://www.instituteforgovernment.org.uk/publications/lifting-lockdown

Lim, J., & Lee, H. C. (2019). Comparisons of service quality perceptions between full service carriers and low cost carriers in airline travel. Https://Doi.Org/10.1080/13683500.2019.1604638, 23(10), 1261–1276. https://doi.org/10.1080/13683500.2019.1604638

Moro, S., Esmerado, J., Ramos, P., & Alturas, B. (2019). Evaluating a guest satisfaction model through data mining. International Journal of Contemporary Hospitality Management, 32(4), 1523–1538. https://doi.org/10.1108/IJCHM-03-2019-0280

Perez, M. (n.d.). Web Scraping vs Data Mining: What's the Difference? | ParseHub. Retrieved January 11, 2021, from https://www.parsehub.com/blog/web-scraping-vs-data-mining/

Preethi, P. G., Uma, V., & Kumar, A. (2015). NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of scientific ScienceDirect Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction. Procedia Computer Science, 48, 84–89. https://doi.org/10.1016/j.procs.2015.04.154

Restuputri, D. P., Indriani, T. R., & Masudin, I. (2021). The effect of logistic service quality on customer satisfaction and loyalty using kansei engineering during the COVID-19 pandemic. Http://Www.Editorialmanager.Com/Cogentbusiness, 8(1), 1906492. https://doi.org/10.1080/23311975.2021.1906492

Sangkaew, N., & Zhu, H. (2020). Understanding Tourists' Experiences at Local Markets in Phuket: An Analysis of TripAdvisor Reviews. Journal of Quality Assurance in Hospitality & Tourism, 1–26. https://doi.org/10.1080/1528008X.2020.1848747

Sheng, J., Amankwah-Amoah, J., Khan, Z., & Wang, X. (2020). COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions. British Journal of Management, 1467-8551.12441. https://doi.org/10.1111/1467-8551.12441

Tripadvisor Help Center. (2021). Retrieved September 27, 2021, from https://www.tripadvisorsupport.com/en-GB/hc/traveler/articles/438

Uğur, N. G., & Akbıyık, A. (2020). Impacts of COVID-19 on global tourism industry: A cross-regional comparison. Tourism Management Perspectives, 36, 100744. https://doi.org/10.1016/j.tmp.2020.100744

UNWTO. (2021). Tourism and COVID-19 – unprecedented economic impacts | UNWTO. https://www.unwto.org/tourism-and-covid-19-unprecedented-economic-impacts

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment Analysis in TripAdvisor. IEEE Intelligent Systems, 32(4), 72–77. https://doi.org/10.1109/MIS.2017.3121555

Attachments

A - Scientific Article

The following attachment contains the submitted article to the ICMarkTech 21 international conference and to be published in the Proceedings by Springer. The following document should be opened with the “Adobe Reader” application in order to preserve the link to the file attached.

Paper_TextMining_
Submission_Camera